

Dangers of publication by press conference

Last week's announcement of an unpublished observation of a possible extra-solar planet raises questions about the wisdom of NASA. The agency's need for visible success could undermine public confidence.

Publication of preliminary results by press release is official policy at the US space agency NASA. That was the message implicit in statements last week by Ed Weiler, head of NASA's 'Origins' programme, at a press conference that issued a photograph of a blob that may be a planet associated with a newly forming star system (see page 406). Unfortunately for those interested in the scientific details, there is only an abstract of a conference submission to turn to. But if NASA wanted publicly and promptly to stamp its ownership on the results, it certainly succeeded.

One does not need to read between lines to perceive a deep need within NASA for publicity. Of course publicity in itself is no bad thing; scientists usually benefit from it, as do those journals — including *Nature* — and organizations that cultivate it. Indeed, NASA deserves much credit for an unsurpassed track record in stimulating public interest in science, combined with an awareness of the need for balance and scientific caution. But there are dangers in the process which NASA now appears to be compounding.

One danger arises paradoxically through the strong attention given to science in the mass media. Scientists may be delighted to find science on the daily news pages, rather than ghetto sections, of newspapers. But it can be a Faustian bargain; the pressures on journalists can on occasion — perhaps often, in under-resourced newspapers — leave them with insufficient time to do much more than turn a press release into something comprehensible and sparkling, possibly excessively so. Investigating a research community's broader perspective is too often prohibited by deadlines.

In that situation, NASA's approach last week, if a harbinger of things to come, is likely to undermine the coverage of science. First, the lack of detail in a readily available publication greatly reduces the ability of journalists who are remote from the action, and of scientists whom they might consult, to do justice to the story. Second, a conse-

quence of that is that the public will more often be faced with sensational and triumphant stories that subsequently prove to have been exaggerated or false — a phenomenon that is itself damaging to perceptions of science.

With its unrivalled capacity to command the attention of a large audience, NASA has all the more responsibility to act scrupulously. It is worrying, then, that NASA bypassed the standard peer-review process in last week's episode. Given this journal's vested interest in such matters, the facts should be allowed to speak for themselves. According to its author, Susan Terebey, the manuscript describing the observations has not yet been completed. (One can only doubt that she, and other astronomers in NASA's hot seat before her, enjoyed the public pressures in supporting the agency's goals.) And according to Weiler last week: "we learned of Susan's observations a couple of weeks ago.... The trouble is, Susan submitted her paper to the American Astronomical Society meeting and that abstract went on the Internet. So it essentially was public as of then.... Before [we went public] we put Susan through a grilling — I would maintain a much more severe grilling than the average paper gets in the average journal.... We had five PhD astronomers sit down with Susan and literally [*sic*] grill her with very tough questions for about an hour and a half... and long discussions with [other scientists participating in the press conference] before we went on air...."

That is indeed a form of peer review. But, in principle, NASA's procedure was not detached — organized as it was by the very institution that funded the research and which has a vested interest in a positive and spectacular result. Both in principle and in its practical implications for journalism, therefore, the rush to release a preliminary result is questionable. If it continues with such an approach, NASA risks undermining the respect for objectivity on which the public support for science ultimately rests. □

Hardening the 'soft' sciences

Environmental sciences are not the only disciplines that would benefit from quantitative stiffening.

'Soft science' is a phrase used by some to mean research that is little more than descriptive, lacking a theoretical and quantitative basis that permits specific predictions that can be tested. Given its indiscriminate use and its disparaging overtones, the phrase is of doubtful value. But a call last week from the head of an environmental research funding agency is one symptom of a progressive hardening of sciences traditionally viewed by some as 'soft' but which are also, as it happens, likely to be critically important for the successful management of the planet.

John Krebs, head of the United Kingdom's Natural Environment Research Council, is surely not the first person to call attention to the need for more mathematicians and physicists to engage in environmental sciences (see page 400). He is no doubt battling against an unjustified but chronic dismissive attitude by some

'hard' scientists towards the disciplines that he funds. He is also fighting an ignorance of opportunities as our ability to understand the Earth's systems through experiment and simulation grows. In that ever more complex context, the ability to focus on physical essentials with confident numeracy becomes correspondingly more valuable.

Krebs also asks molecular biologists to help in the protection and sustainable use of the environment. That, too, is a timely summons, but there is also a separate crusade to be waged with them. As physics, chemistry and mathematics offer more in understanding the behaviour of biomolecules and systems in which they interact, so there is a need for a quantitative strengthening of biologists' training. Is that self-evident, and are universities responding, or is there chronic resistance there too? □



'Rough draft' of human genome wins researchers' backing

[WASHINGTON] The major players in publicly funded efforts to sequence the human genome are proposing to create during the next three years a 'rough draft' of the genome that would be about 95 per cent complete.

Such an initiative would add a new element to a strategic plan that is being drawn up to guide the Human Genome Project (HGP) in the next five years. The sequencers say a 'rough draft' would allow other scientists to proceed more rapidly with projects that apply new-found sequence data, from discovering rare disease genes to pinning down the molecular details of disease genes that have been mapped but not yet identified.

It would also be different to the 'whole genome shotgun' sequencing approach recently announced by J. Craig Venter and the equipment manufacturer Perkin-Elmer, (see *Nature* 393, 101, 201 & 296; 1998).

"From the point of view of identifying genes, [the 'rough draft'] would be absolutely fantastic," said Leroy Hood of the the University of Washington in Seattle during a two-day meeting in Virginia last week to refine details of the five-year plan.

The new strategy became a focal point of the meeting, with participants endorsing the idea that experiments should begin to see how feasible and accurate it would be. Francis Collins, the director of the National Human Genome Research Institute (NHGRI), said the strategy would speed up the process of getting "very useful" sequence "in the hands of the people who want it".

But Collins, whose institute is the main government funder of the genome project and also sponsored the meeting, emphasized that the rough draft "is not a substitute for finishing the whole thing".

Federally funded sequencers are using a two-part process to sequence the genome by analysing mapped clones covering the human chromosomes. The first step is 'shotgun' sequencing and assembly of random fragments from each clone; the second is the tedious and expensive process of 'finishing' by closing gaps and resolving uncertainties.

The new proposal involves dramatically accelerating the shotgun phase, which is the simplest and cheapest component (presently only 10 cents per base). Researchers estimate that this will require a two- to two-and-a-half-fold increase in the government's sequencing capacity, and could produce a high-quality rough draft by 2001. The complete version should still be finished by 2005.

The approach of Venter and Perkin-Elmer breaks the genome into random unmapped fragments for sequencing. But some scientists doubt whether the whole human genome, 70 per cent of which consists of highly repetitive sequences, can be reassembled.

The government's approach requires assembly of much smaller regions, each about one twenty-thousandth of the genome, decreasing the possibilities for error by a factor of 400 million.

The participants at the meeting said their project would mesh well with the private

venture by providing accurately assembled sequence across the genome, which the private venture's own sequence data could fill out. "It may be that the two together will be adequate to finish the job. If that's the case, everybody wins," said one sequencer.

But despite excitement at the prospect of producing a useful 'intermediate product' in three years, some sequencers thought the project could distract them from the primary goal of producing a complete, highly accurate, finished human sequence by 2005. The sequence-ready maps needed for shotgun sequencing of clones do not yet exist across the entire genome, and their construction could distract sequencing centres that are already stretched to meet the 2005 goal.

Nor is it clear whether Congress will allocate sufficient funds to double sequencing capacity quickly, even though government sequencers have said their labs could handle the extra capacity.

But nearly all participants agreed that the costs of raising sequencing capacity are modest compared with the long-term stakes and that more government sequencing capacity is needed urgently.

"More would be better, sooner rather than later," said Richard Lifton, a Howard Hughes Medical Institute investigator at the Yale University School of Medicine, who co-chaired the sequencing strategy session.

Collins says that the federal project is "massively undercapitalized" with regard to sequencing capacity. The consequences of not responding to that, he said, are not only that the finished human sequence would not be produced quickly enough, but that medical advances and the sequencing of other animals would be delayed.

He also says a push for faster sequencing makes sense on economic grounds alone: a "very rough" estimate of the current annual costs of US public and private cloning and sequencing efforts is \$500 million.

The draft five-year-plan circulated by NHGRI to participants — which will be modified by their input and released in final form in October — calls for the generation of 500 million base pairs of sequence per year by 2003, a tenfold increase on current levels.

Participants said their focus on the shotgun strategy, and boosting speed and capacity generally, was not influenced by the Perkin-Elmer/Venter venture. Rather than federal plans being "ignited" by the private effort, said Lifton, "the same opportunities that led to this new enterprise also led to realization within the Genome Project that there was a real opportunity to accelerate" sequencing.

Very Large Telescope reveals first images



[LONDON] The European Southern Observatory last week unveiled the first images from its Very Large Telescope on Mount Paranal in Chile. The colour picture (left) shows the fine structure of the Butterfly nebula. It is a composite of three exposures through broad-band blue, green and red filters.

The optical/infrared telescope's image quality has surpassed expectations. The observatory's staff say the angular resolution is unequalled by any ground-based telescope. Its sensitivity in resolving point sources — such as stars — should be superior to anything achieved by a telescope on Earth.

As if to show that the public and private efforts are cooperating, Michael Hunkapiller, president of Perkin-Elmer's Applied Biosciences Division and its top official on the private venture, and Mark Adams, who works with Venter at The Institute for Genomic Research in Rockville, Maryland, were at the conference.

Also present were investigators from the federally funded sequencing laboratories, other university scientists, and officials from NIH and the Department of Energy, the other major government funder of the HGP, along with representatives of Britain's Wellcome Trust, which recently agreed to fund one third of the total sequencing effort of the human genome (see *Nature* 393, 201; 1998).

Hunkapiller fielded pointed questions about Perkin-Elmer's intellectual property intentions, and one federal sequencer predicted that the product of the new venture would be riddled with errors, but the overall tone of the gathering was collegial.

Harold Varmus, director of the National Institutes of Health (NIH), declared his pleasure at the private-sector participation, and urged federal scientists to make the public and private efforts "consensual, communal, collaborative and productive". He added: "There's every reason to believe that we can do that." Hunkapiller said that for the two efforts not to collaborate is "absurd".

The private group promised prompt and full public access to its raw sequence data, but publicly funded scientists and federal officials remain wary. Collins said that, although the venture's leadership was "reassuring" on the question of public access, he worries that "down the road somebody else who ends up as calling the shots may see this as too much of a give-away and retreat". **Meredith Wadman**

UK seeks physicists for environmental research

[LONDON] Britain's Natural Environment Research Council (NERC) is to introduce a series of fellowships designed to encourage more physical scientists, computer scientists and molecular biologists to take up research on the environment.

"The environment presents the largest, the most serious and the most intractable challenges for the next 50 years," says John Krebs, NERC's chief executive. He says this requires a "new breed" of scientist, and new ways of problem solving that cut across traditional disciplines.

The main goal of the fellowships is to tackle an expected shortage in Britain of environmental scientists with mathematical, computational and statistical skills. A second aim is to involve more natural and social scientists in helping to solve questions in environmental research.

NERC hopes to get molecular biologists to work with ecologists in prospecting for genes from plants with medicinal potential, for example, and in using biotechnology to clean up pollution in soil and water.

The council also wants to see more social scientists working with Earth scientists in research that places values on environmental goods and services, and on the planning and management of towns and cities.

Up to 30 four-year grants could be on offer, paying around £50,000 (US\$82,000) a year. The sum includes the salaries of the main postdoctoral researcher and a research assistant, as well as research support.

"The big questions on the environment cannot be solved by one or two narrow disciplines on their own," says Krebs. "Many of the most innovative advances occur at the interfaces between disciplines."

Krebs says that NERC is not downgrading the importance of basic research, but plans to increase its investment in curiosity-driven research. It will soon launch five-year response-mode grants for basic research projects; its existing grants last three years.

"Anecdotal evidence tells us that there are not enough people with a physics or mathematics background going into the environmental sciences," says Krebs. "We need to bring such people to help tackle problems in meteorology, hydrology and in the oceans."

The grants and fellowships are expected to be announced after the government's Comprehensive Spending Review, due next month. Both schemes are part of NERC's latest science strategy, which was unveiled at the end of last month. A major aim of the strategy is to remove some of the uncertainty of research careers by tackling the issue of short-term contracts.

NERC's proposed four-year fellowships and five-year basic research grants are designed to provide more stability in research careers, and to encourage long-term interest in a field, says Krebs.

All contract staff employed for more than five years at NERC's research centres and surveys will either be offered permanent posts or released. **Ehsan Masood**

Australian advisory body's first target is to reduce land salinity

[SYDNEY] At its first meeting since being restructured, the Australian Prime Minister's Science Engineering and Innovation Council has decided its most immediate task is to find solutions for the rising salinity which is adversely affecting the meagre soils supporting the nation's agriculture.

After more than a century of bush clearance, deforestation and irrigation, salinity has become a major threat to income from exports and the quality of the environment. John Stocker, the chief scientist, estimates that salinity now affects about 5 per cent of the land sown to crops or pastures, and contributes 12 per cent of the yield lost to land degradation.

The state of Western Australia is losing 200 hectares from its wheat belt every day and, with salinization increasing and proving difficult to treat or reverse, salt-affected land is expected to increase

sevenfold to 20 million hectares by 2050.

The move followed the acceptance of a scheme by Stocker for rationally determining preferences for the national support of research and development, which has been an unrealized goal of successive Australian governments. Stocker defines five criteria, which he calls "structural priorities", for assessing the public funding of science and technology.

Stocker argues that research programmes should be gauged by their effects on maintaining the national "science base", developing "applicable knowledge", promoting "interaction among providers and users" of research, stimulating "innovation in industry", and improving "awareness" of science and technology.

The council has expanded by absorbing the role of the Australian Science and Technology Council (see *Nature* 391, 624; 1998). Representatives of science

organizations have been augmented with leaders of business and grant-giving bodies, and should provide ministers with "high-level and independent advice", said prime minister John Howard in a statement.

The council told a working party, led by Stocker, to produce by August an action plan on salinity. "We'll have to demonstrate some agility to do it by then," Stocker says. His report may become caught in the last weeks of an (as yet undeclared) election campaign, in which Howard and Kim Beazley, the leader of the Labor opposition, are trading blows about tax reform.

Indications that science is unlikely to figure prominently in the election are reinforced by Howard's not capitalizing on the council's positive outcome to make a political point. But Beazley has called for urgent correction of the Coalition's declining taxation incentives for R&D in industry. **Peter Pockley**

Pakistan weighs up impact of sanctions...

[LONDON] Pakistan's scientists are bracing themselves for further restrictions on their access to science and technology facilities in Western countries following their government's decision to carry out nuclear tests last week.

But some believe that such restrictions could benefit science in Pakistan by catalysing collaboration between Muslim countries. That is the view of Atta-ur-Rahman, a Pakistani chemist who is coordinator general of Comstech, the Islamabad-based agency responsible for scientific cooperation between the 50-member Organization of Islamic Countries (OIC).

Pakistan is already prohibited from acquiring defence-related technologies from the United Kingdom and the United States. Further restrictions are widely believed to be imminent following the testing of six nuclear devices last week.

One target could be the Abdus Salam International Centre for Theoretical Physics in Trieste, Italy. The centre is a United Nations organization controlled by the International Atomic Energy Agency (IAEA). It was set up by Pakistan's Nobel prize-winning physicist Abdus Salam to enable scientists from developing countries to collaborate with colleagues from the developed world.

Miguel Virasoro, the centre's director, says he expects an "intelligent reaction" — possibly in the shape of restrictions to some of the centre's programmes related to nuclear energy. But he says it is unlikely that Indian and Pakistani physicists will be com-

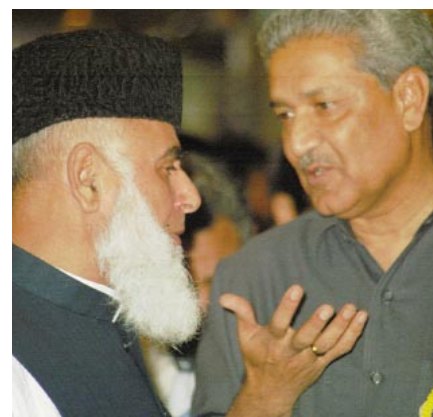
pletely isolated. Most of the centre's programmes are "simply basic research" in areas such as high-energy and condensed matter physics, "not nuclear physics", says Virasoro. "However, if the IAEA decides to take action against Pakistan's scientists, we will have to follow that policy" he adds.

Existing technology restrictions on Pakistan include curbs on importing equipment — such as supercomputers — and prohibitions on students doing research in areas such as nuclear physics and lasers.

In contrast to the reaction in the West, Pakistan's nuclear tests have generated overwhelming popular support in Muslim countries. Muslim governments have not come out in support for the tests. But newspaper articles in much of the Middle East have hailed the emergence of the first Muslim nuclear weapons state.

Many of Pakistan's scientists feel their country should take advantage of its new-found status. "The nuclear tests will certainly raise the image of Pakistan as a leader in the field of science and technology in Muslim countries," says Rahman. "This can certainly strengthen the role of Comstech."

Although Comstech was promised US\$10 million last year by OIC member states for various projects, it has received less than half that amount. But it was recently awarded \$2.1 million by the Jeddah-based Islamic Development Bank for scholarships for 190 students from the 18 least developed member states to train and do research within the OIC's leading institutions.



A. Q. Khan (right), founder of Pakistan's nuclear programme, with President Rafiq Tarar.

Abdul Qadeer Khan, founder of Pakistan's nuclear programme and president of the Pakistan Academy of Sciences, is another keen advocate of stronger scientific links between Muslim countries. Khan asks why technological restrictions apply only to certain Muslim states and not other nations. He has in the past supported a call for the setting up of an atomic energy commission for Muslim countries, and stronger links between defence research agencies in the Muslim world.

But such an approach is not without its critics. Iftekhar Ahmed, a Pakistan-born experimental physicist who is president of the Commonwealth Academy of Sciences in London, is deeply critical of the decision to explode nuclear devices, and believes there is little to be gained scientifically by collaborating with Muslim countries.

Ahmed says virtually all of Pakistan's leading scientists were trained, not in Muslim countries, but in Europe or the United States "because that is where the best science is done". He warns that Pakistan is risking its future research base by losing access to Western facilities.

"Before, a fledgling person like me who came from the back *galis* [alleys] of Lahore could train at the highest levels and work with the best minds," he says. "Now there will be no more Salams. And even no more Qadeer Khans. We have starved our future generations."

Virasoro, too, questions the scientific benefits of closer ties to Muslim countries when most advanced technology is in the developed world. He says nuclear weapons represent "yesterday's technology", which does not need much research, only resources that could be better spent elsewhere.

"Pakistan is deluding itself if it thinks that exploding a bomb turns it into an advanced, high-technology country," he says. "It is folly to think that you are more advanced than Sweden or Norway just because you have the bomb [and they do not]."

Ehsan Masood

...as India sees decline in cooperative work

[NEW DELHI] Indian scientists fear that the subcontinent's 'tit-for-tat' nuclear tests are likely to set back significantly recent attempts to improve cooperation with colleagues from Pakistan.

The prospect of a 'cold war' between the countries is expected to raise the political pressure on scientists from both sides to reduce contact with each other. One event whose fate hangs in the balance, for example, is a conference for scientists from non-aligned countries that Pakistan had proposed to host this year.

"Only last week we had received a letter from Pakistan offering to host the conference," says K. N. Johry, director of the Centre for

Science and Technology of the Non-Aligned Countries, whose headquarters is in New Delhi. "We do not know if the offer still stands."

Pakistan has been an active member of the centre, which promotes scientific exchanges among its 38 member nations. "Scientists from Pakistan played key roles in meetings on medicinal plants organized by the centre last December, and another on DNA fingerprinting," says Johry.

Another potential casualty is the year-old Indo-Pakistan Friendship Forum, which aims to bring Indian and Pakistani health experts together to tackle common problems such as maternal mortality and hepatitis B.

The project's co-founder, Vulimiri Ramalingaswami, a leading health scientist in the region, says it is "lying dormant now". But he hopes to revive it "once the nuclear dust settles".

Indira Nath, former foreign secretary of the Indian National Science Academy in New Delhi, says the political fall-out from a potential nuclear arms race will set back scientific exchanges.

The Indian and Pakistani science academies do not have direct exchanges "but we get involved with each other through the Federation of Asian Science Academies," says Valangiman Ramamurthi, secretary of India's science ministry.

K. S. Jayaraman

Plea for female academics in Germany

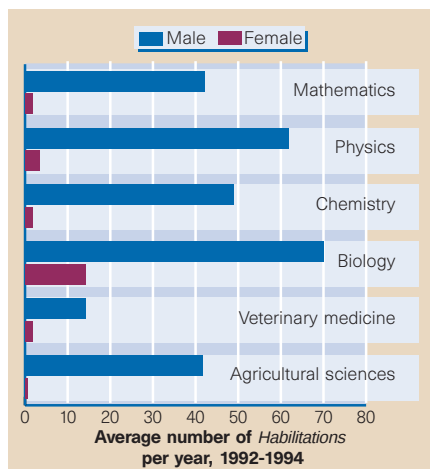
[MUNICH] German universities should introduce recruitment and employment policies designed to increase the number of female academics, and should be rewarded financially for their success in doing so, says a report by the Wissenschaftsrat, Germany's influential science council.

The report was drawn up by a panel chaired by Dagmar Schipanski, a former president of the Wissenschaftsrat. One suggestion is that German universities should move towards the Anglo-Saxon system of postdoctoral fellowships, which is more attuned to the family demands on women.

The German academic community is one of the most male-dominated in the industrial world. Twice as many men as women complete PhDs — in physics the ratio is as high as 13 to one — and only 4.5 per cent of those in top academic positions are women.

In its report, the Wissenschaftsrat says that Germany is well placed to influence these ratios in the near future, as many faculty members who were hired during the university expansion in the 1960s and 1970s are due to retire in the next few years. But it says that a major restructuring of the academic system is needed to achieve this goal.

For example, German women who wish to raise a family are often discouraged from choosing an academic career by the length of time needed to gain the qualifications required for a university post. These include not just a PhD, but also *Habilitation*, an advanced qualification designed to prove research and teaching skills (see chart). *Habilitation* candidates are attached for many years to a particular professor, and the average age of qualification is around 40.



The Wissenschaftsrat's report, entitled *Recommendations on Equality of Opportunities for Women in Science and Research*, supports a many-pronged approach. It says that universities should take steps to encourage more women to complete PhDs, for example by limiting PhD duration to three years. It also suggests that universities should ensure that the proportion of women in academic posts — including departmental chairs — is at least as high as the proportion gaining academic qualifications in the same subject.

Rather than enforcing quotas, the Wissenschaftsrat suggests rewarding universities when targets for employing women are met. It suggests that success in promoting women should be a criterion in allocating funds to universities, an idea recently put into practice in the west German state of Hessen.

The report warns against relying on special programmes designed to support female

PhD students and *Habilitation* candidates without a new hiring policy being in place.

The report issues a somewhat circum-spect attack on the system of *Habilitation*. It says that, if a system of postdoctoral fellowships were introduced, research and teaching experience gained in such positions should be accepted as an alternative to *Habilitation* as proof of competence. *Habilitation* is often criticized as unnecessary and restricting (see *Nature* 384, 305; 1996), but is staunchly defended by academic conservatives, who fear the loss of local power for professors.

As the long-drawn-out training means that academics may first apply for permanent posts only around the age of 40, the report recommends that universities should abandon their system of *Hausberufsverbot*, whereby an academic position may not be given to a candidate from the same university. Intended to encourage cross-fertilization among universities, this often works against women with young families who are unable to move cities at that stage in their lives.

Experience at other universities should be gained at an earlier stage, says the report. Universities should advertise PhD and post-doctoral positions nationally to encourage mobility among the young. The report points out that this would also increase the quality of competition for such posts.

The report makes the oft-repeated call for more flexible child-care facilities in universities, and for men to accept their share of family responsibilities.

Hans-Gerhard Husung, a spokesman for the Wissenschaftsrat, says that the short-term and mid-term recommendations form a package that could make a serious difference to an urgent problem.

Brigitte Mühlenbruch, spokeswoman of the German Conference for Quality for Women in Universities, agrees. If all the recommendations are approved, she says, "they will make a big difference to women's employment prospects".

Dagmar Schipanski says she regrets the fall in opportunities for female academics in east Germany since reunification. A much higher number of female than male academics lost their jobs as a result of reunification, and the generous child-care facilities at universities in the east were heavily cut back.

Women were present at all levels of academic life in a higher proportion in the old East Germany than in the west. But Schipanski points out that, despite adequate child-care facilities, the proportion of women at higher levels of the career pyramid still decreased in a manner similar to that in the west.

The Wissenschaftsrat's recommendations do not have legal authority, and it is up to universities to decide if and how far they are adopted.

Alison Abbott

Authors slow to retract 'fraudulent' papers

[MUNICH] One year after Germany's biggest fraud scandal became public (see *Nature* 387, 442; 1997), only two of the 47 scientific papers tainted have been retracted, says the German daily *Süddeutsche Zeitung*.

According to the newspaper, some co-authors are reluctant to allow retraction out of fear that their careers may be damaged if their names become associated with the fraud. Some young co-authors are said to have threatened to sue the committees investigating the fraud, rather than supporting them by retracting papers.

The *Süddeutsche Zeitung* asked the 19 journals that published most of the papers believed by investigating committees to be fraudulent how they had handled the case. The two papers that have been retracted were in the *Proceedings of the National Academy of Sciences* and in *EMBO Journal*. According to the newspaper's survey, many

of the other journals, including the *Journal of Biological Chemistry*, the *Journal of Immunology* and *FEBS Letters*, claimed that they were completely unaware of the affair.

In seven cases Friedhelm Herrmann, professor of medicine at the University of Ulm — one of the two researchers accused of the fraud, and corresponding author of most of the publications — refused the requests of journal editors to retract papers.

Herrmann says he was unaware that fabrication of data was taking place in his laboratory. Much of the contested work took place in the laboratory headed by Herrmann at the Max Delbrück Centre for Molecular Medicine in Berlin in the early 1990s.

Scientific journals have different policies on the retraction of papers. Some require only the consent of the corresponding author, while others require the agreement of every co-author.

Quirin Schiermeier

NASA moves to speed up grants process

[WASHINGTON] The US space agency NASA has begun to streamline its grants administration process after a review found widespread dissatisfaction among scientists who depend on the agency for research funding.

A series of reforms still awaiting final approval by top managers is intended to cut the lag between winning a grant and receiving funds to less than a month within two years. The wait for money at present is between 77 and 167 days, according to NASA.

A team led by Mary Kicza, associate director for space science programmes at the Goddard Space Flight Center in Maryland, began looking at problems with the management of grants last summer in response to complaints from the research community

which caught the attention of Congress and the White House (see *Nature* **389**, 218; 1997).

To understanding the scale of the problem and its possible solutions, Kicza's group sent letters to more than 6,000 scientists, consulted advisory groups, held open meetings, and met grant managers at other federal agencies.

Typical complaints from NASA-funded scientists included late payment, lack of predictability in the grants process, and burdensome paperwork. The solutions, to be fully implemented by 2000, would involve switching to more efficient electronic grants tracking, standardizing forms and deadlines for NASA Research Announcements (NRAs), and reducing the number of sign-offs needed

for grant solicitation and approval.

Even before the plan is fully in place, the agency has started reducing the number of signatures required for some steps in the process, Kicza says. Extra staff have been brought in to help work through the backlog.

Switching to electronic grants processing will also speed things up, but will be introduced more slowly as part of an agency-wide overhaul of accounting practices called the Integrated Financial Management Project.

The proposed goals are ambitious. The lag from grant selection to award of funds is targeted to drop to 62 days by the end of this year, to 46 days in 1999, and finally to 28 days in 2000. The agency hopes to shorten the interval between releasing an NRA and selecting winners from 285 to 240 days.

It hopes eventually to process annual grant renewals in only 16 days, down from the current 79 to 95 days. Principal investigators whose funding requests change by 20 per cent or less from one year to the next will not be asked to file new proposals. Grant applicants will also be able to track the progress of their proposals via the Internet.

Duplicative paperwork and too many people in the approval queue cause most delays, according to the Kicza team, which advocates a mostly paperless process. Grant information now is keyed in several times into separate and mutually incompatible databases, but the goal is to key it in once to an integrated electronic accounting system.

The reaction from space scientists is very positive, says Kicza. David Black, director of the Lunar and Planetary Institute in Houston, who heads a separate review of NASA research grants, says: "It's a tough goal [NASA] has set for themselves, but I think the only way you're going to really make the system respond is to push real hard."

David Burrows, a Pennsylvania State University astronomer and one of several scientists asked to review the new grants procedures, agrees that the targets are ambitious. But he says: "I think they can do it if they put their mind to it" and if they have support from top NASA managers.

But one temporary block is staffing within the agency's Office of Space Science. Last week the outgoing director of the its research programme, Henry Brinton, sent a plaintive letter to the space science community encouraging applicants for two positions that have remained unfilled for weeks. The posts are for 'discipline scientists' and include administering NRAs.

"The Office of Space Science is disappointed at the lack of positive response from the scientific community," wrote Brinton in his letter. As a result, "it is increasingly likely that both grants programs will suffer significant delays".

Tony Reichhardt

South Africa gives green light to telescope



Building sight: artist's impression of SALT beside existing telescopes at Sutherland in South Africa.

[CAPE TOWN] Lionel Mtshali, South Africa's minister of Arts, Culture, Science and Technology, announced this week that the government is to foot half of the bill for the construction of the new Southern African Large Telescope (SALT).

The nine-metre instrument for optical and infrared astronomy will be the largest single optical telescope in the Southern Hemisphere. The government will contribute 50 million rand (US\$10 million) towards the cost, its largest investment so far in a single scientific project.

"South Africa is sending a strong signal to the world that we are prepared to support fundamental science at an international level," says Khotso Mokhele, president of the Foundation for Research Development, the research council under whose aegis the observatory falls.

The telescope, to be erected at the existing Sutherland site of the Southern African Astronomical Observatory (SAAO)

in the Northern Cape Province, will complement the Hobby-Eberly Telescope at the McDonald Observatory in Texas (see *Nature* **384**, 13; 1996).

The balance of the costs may be met by foreign sources in exchange for observing time. "South Africa cannot go it alone in a project of this magnitude," says Bob Stobie, SAAO director and chief architect of the project, who says he is "naturally delighted" with the government support.

Patricia Whitelock of SAAO says SALT's location and excellent viewing conditions will help studies in four areas: the early Universe and galaxy formation; galaxy populations; planetary searches; and quasars and active galactic nuclei.

A feasibility study has shown that South African industry can complete about half of the project's construction requirements. Technology transfer with US companies could have spin-offs for related areas of South African industry.

Michael Cherry

Storm in Japan over sale of zoo monkeys for research

[TOKYO] The Japan Monkey Center, a primate research centre in Inuyama, Aichi Prefecture, has come under attack from animal rights activists for arranging the sales of monkeys obtained from zoological parks to medical research laboratories.

More than 160 Japanese macaques were obtained from zoos across Japan this year to be sold to research institutes for ¥75,000 (US\$540) each.

Animal rights groups, such as All Life in a Viable Environment (ALIVE), last week criticized the centre's role as unethical, saying that zoo animals should not be sold, let alone be "savaged" in animal experiments.

But the monkey centre is defending its position, saying that the macaques were obtained from zoos suffering from an overpopulation of the animal, and that experiments using primates are crucial for the advance of medical sciences.

The shortage of research monkeys is a serious concern for biomedical researchers in Japan. Demand for the animals is rising with advances in brain sciences and research on HIV vaccines and gene therapy, and many scientists say that the high cost and shortage of supply is starting to affect their work.

Last year, the Japan Neuroscience Society submitted a report to the Ministry of Education, Science, Sports and Culture, raising problems generated by the lack of research monkeys in Japan.



Twist of fate: the sale of Japanese macaques has provoked protests from animal activists.

According to the report, the demand for research monkeys has increased by 100 per cent in ten years, and most scientists are dependent on animals obtained from the wild. Although some national institutes, such as Kyoto University's Primate Research Institute, breed several hundred monkeys each year, external researchers do not have access to such animals.

The shortage of monkeys is partly the result of a decision made eight years ago by Japanese airlines to stop importing monkeys from Africa and South-East Asia to prevent the entry of the Ebola virus (see *Nature* 382, 744; 1996). Restricted supply has caused the price of monkeys to increase, and many scientists have been forced to use different species of monkeys.

Asako Saegusa

France seeks the return of the roving scholar

[PARIS] Lionel Jospin, the French prime minister, last week urged the creation of "a university without walls" in Europe, similar to that which allowed the free flow of intellectuals in Europe during the Renaissance.

Jospin was speaking at the end of a two-day colloquium held at the Sorbonne in Paris, marking the 800th anniversary of the founding of the University of Paris — though the date is contested by some. He said the goal should be "for a student to be able to start his studies in Paris, and follow them at Heidelberg, Oxford or Bologna". He added: "If the rules get in the way of that mobility, then we need to change the rules."

At the end of the meeting, education ministers from the United Kingdom, France, Germany and Italy signed a declaration committing themselves to building a common European university system. This would include harmonizing course structures and encouraging greater mobility of students and staff among European countries.

Jospin, himself a former minister of education, said that a new political impetus was needed. He endorsed the declaration's statement that "Europe is not only that of the Euro, of the banks and the economy; it must be a Europe of knowledge as well".

The four ministers — Claude Allègre (France), Tessa Blackstone (United Kingdom), Jürgen Rüttgers (Germany) and Luigi Berlinguer (Italy) — committed themselves at the meeting to harmonizing their university degree systems. They agreed that this should cover a first degree, a short masters and a longer doctorate degree, including a system for transferring credits obtained for modules and semesters in various European institutions.

This year, around 200,000 students and 31,000 lecturers will take part in the European Commission's exchange schemes such as Socrates and Leonardo. But this is a fraction of the continent's 11 million students.

Discouraging factors include excessive bureaucracy and the relatively low value of grants within the schemes compared with the costs of living abroad. But the lack of mutual recognition of course credits is generally seen as an even larger obstacle.

The ministers backed credit transfer schemes, such as the commission's recently launched European Credit Transfer System, to which around 1,300 universities have signed up.

But the consensus at the meeting was that a new political effort was needed, in particular to take advantage of the opportunities offered by new technologies such as video-conferencing.

Declan Butler

Lawsuit demands labels for modified foods

[WASHINGTON] A group of US consumer advocates and scientists is suing the Food and Drug Administration (FDA) for failing to require manufacturers to label genetically modified foods as such, and for what they see as shortcomings in the regulation of their safety.

In a suit filed at the US District Court in Washington DC last week, lawyers at the International Center for Technology Assessment (ICTA) claim that the FDA is violating the Federal Food Drug and Cosmetic Act by failing to require the labelling of the foods, which are now being widely grown and eaten in the United States.

The lawsuit also argues that the FDA should be treating genetic modifications as new food additives, which need to be tested for safety and approved before being sold.

The FDA says that it is not required to label genetically modified foods because they are not materially different from existing foodstuffs. "If the food is different, we require labelling," says Eric Flamm, a

policy analyst at the agency. "If it isn't different, we do not."

The agency's biotechnology policy, published in 1992, states that genetic modifications are not regarded as new additives for regulatory purposes, provided they use genes taken from items that are already in the food chain. The lawsuit claims that the agency should be treating all genetic modifications as additives. It challenges what it calls the FDA's failure to label or to regulate 36 genetically altered foods, including corn and tomatoes.

"By failing to require labelling, the FDA has made millions of Americans into guinea pigs for genetically engineered foods," says Andrew Kimbrell, an ICTA lawyer.

The plaintiffs include Philip Regal, professor of ecology at the University of Minnesota, who claims that the FDA is too close to the industries it is supposed to regulate. Regal says that safety risks from genetically engineered foods are "a scientific certainty".

Colin Macilwain

Swiss transgenic vote hangs in balance

[MUNICH] Tension is running high among scientists in Switzerland on the eve of the country's referendum on genetic engineering, to be held next Sunday (7 June).

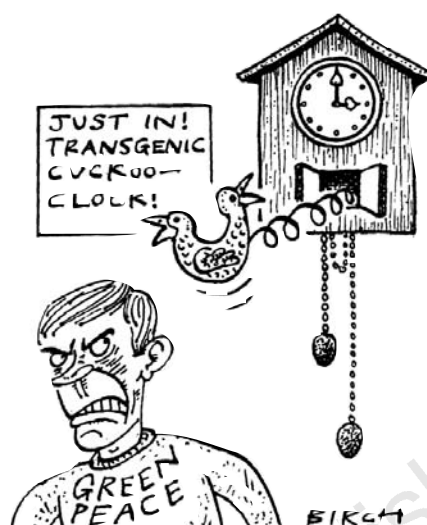
Some are backing an initiative provocatively called 'Genepeace', launched by student supporters of genetic engineering in Zürich, Basel and Bern. Genepeace aims to promote peaceful dialogue between geneticists and opponents of genetic engineering — but also warns of the consequences of its possible abandonment in Switzerland.

The referendum proposes a ban on research involving genetically modified animals and on field trials with genetically modified plants (see *Nature* 392, 741; 1998). Despite discreet lobbying by the pharmaceutical industry — and grim warnings of the possible consequences of a 'yes' vote — the outcome of the referendum remains highly uncertain.

The most recent survey, by the German-speaking Swiss Television TVDRS two weeks ago, found that 39 per cent of interviewees were opposed to the ban, 37 per cent were in favour, and the rest were undecided.

But in an earlier survey by the newspaper *Sonntagsblick*, only 16 per cent said they would vote for the abandonment of genetic research, with 32 per cent saying they intended to vote against, while more than 50 per cent were still undecided.

In fact, the result may have already been



decided, because a large proportion of voters has delivered postal votes. Conrad Engler, head of gene technology at Interpharma, the association of the Swiss pharmaceutical industry, says that campaigners for and against the ban are now usually told by the public they approach: "You don't have to convince me any more; I have already voted."

Nevertheless pressure groups have been stepping up their campaigning efforts — some serious, others less so — over the past few weeks. Greenpeace and the World Wild Fund for Nature, for example, are suing the University of Geneva for selling transgenic

mice as animal food to zoos.

Greenpeace has also sued academics — including Klaus Ammann, head of the botanical garden in Bern, and Beda Stadler, a professor of immunology in Bern, both of whom had joined the 'Genepeace initiative' — for abuse of its name and logo.

Genepeace includes among its members Rolf Zinkernagel, director of the Institute for Experimental Immunology at the University of Zürich and 1996 Nobel laureate.

The initiative was launched after 20 or so Greenpeace activists mixed with the crowd of 3,000 students, academics, medical doctors and researchers who demonstrating in April in Zürich against the proposed bans on genetic engineering.

The Greenpeace activists carried placards that differed only subtly from those of the scientists. Greenpeace changed 'Gen Suisse', the name of a foundation supporting gene technology, into 'Gen *Bschiss*', a German word for 'swindle'. In revenge, supporters of genetic engineering found a pun of their own, and launched 'Genepeace'. Ammann and Stadler are among those who have worn T-shirts bearing the 'Genepeace' logo, which, according to Greenpeace, was deliberately intended to confuse the public.

Despite the high emotions that the campaign has generated, little more than 40 per cent of the Swiss population are expected to cast a vote.

Quirin Schiermeier

Italian and US backers improve prospects for cheap trip to Mars

[MUNICH] The lifetime of Mars Express, the fast and cheap voyage to Mars proposed by the European Space Agency (ESA), could be doubled following an offer from NASA and the Italian Space Agency (ASI) to provide the mission's telecommunications system and other assistance.

Under a suggested three-way deal, ASI will build a 64-metre radioastronomy dish, while NASA's equipment would allow the voyager to relay data from the US space agency's Mars Lander in 2005. The Italian agency has also offered to consider providing ground support for the mission. The cost of ground support would increase significantly since the proposed agreement would extend the mission's life from two to four years.

The politically popular Mars Express mission, which it is hoped to launch in 2003, comes up for formal approval by Europe's research ministers in November. Last week's meeting of the agency's Science Programme Committee (SPC) accepted the trilateral agreement in principle. The deal could be ready for signing as soon as the mission is formally approved.

By cutting the overall costs of both ESA's and NASA's Mars missions, the agreement will allow the scientific returns to be increased, says the SPC. It would also open the door to closer coordination of the two agencies' Mars missions, something that both sides would like to see.

Giovanni Bignami, ASI's director of space science, says that the Italian end of the telecommunications package would be paid for by ASI's technology programme, rather than its science programme, so there is no danger that it will impact on ASI's funding of other ESA science projects such as the cometary mission Rosetta.

The telecommunications system would transmit and receive data from experiments and landers on the planet surface from the low orbit of Mars Express. So far, none of the 12 proposed landers for Mars Express has been approved by ESA, as none has secured the necessary assurances of national financing.

But the SPC approved a payload of seven on-board experiments including subsurface radio-imaging of the planet surface's structure, photographic and spectrometric

imaging of the surface, atmospheric physics and magnetospheric physics.

ASI has offered the data collecting services of its radiotelescope currently being built in Sardinia, and which is partially funded by the European Union. This radiotelescope, scheduled to become operational in 2001, will be able to receive not only natural radioactivity from space but also data from planetary probes including Mars Express.

At last week's meeting, the SPC also overruled a proposal from ESA officials to merge two of the agency's planned missions, the 'cornerstone' Far Infrared Space Telescope and the cosmic microwave background surveyor Planck (see *Nature* 387, 639; 1997).

Although a merger would have saved a considerable amount of money, the SPC argued that the scientific return would have diminished to an unacceptable level.

The two missions will now share the same satellite, but will be physically independent. The proposed launch date for the two missions will now slip back one year, to 2007.

Alison Abbott

Brookhaven reactor decision delayed another six months

[WASHINGTON] Prospects for the reopening of a troubled research reactor at the Brookhaven National Laboratory in New York State have receded yet further, raising the prospect that it may never reopen.

The Department of Energy last week told local representatives that the decision about whether to reopen the High Flux Beam Reactor has been pushed back from December this year to May 1999. The problem dates back to January 1997 when the reactor was closed following leakage of tritium from a waste storage tank (see *Nature* 386, 3; 1997).

Local politicians have strongly opposed the reopening of the reactor, which was an important source of neutrons for researchers in the north-eastern United States.

The department says it has received almost 600 comments on an assessment it is preparing on the environmental impact of the reactor, and needs time to digest them. It also said last week that it would comply with a demand from Congress and bring in an outside regulator — the Nuclear Regulatory Commission — to “conduct a comprehensive safety and compliance review” of the reactor.

OECD counts costs of CO₂ emission controls

[LONDON] Stabilizing carbon dioxide emissions at their present levels could cost up to 2 per cent of the world's gross domestic product, according to a report from the Organization for Economic Cooperation and Development (OECD). On the other hand, no action to reduce emissions will cause them to increase three-fold by 2050.

The figures are contained in the OECD's latest *Economic Outlook*. The report calls for developing countries to reduce their emissions, but acknowledges that a way has to be found to divide the costs of environmental protection between the wealthier polluters and developing countries.

Spanish telescope attracts interest

[MUNICH] Following the decision by the Spanish government to finally commit itself to funding a ten-metre optical telescope in the Canary Islands (see *Nature* 392, 852; 1998), the project has now received a number of formal requests to participate.

Potential partners include the United Kingdom, the United States, Italy, Mexico and India, whose joint contributions would cover 35 per cent of total costs. This is the maximum level of participation permitted in

the programme, intended to bring Spanish astronomy up to international levels.

Vanderbilt settles in anaemia cases

[WASHINGTON] Vanderbilt University, in Nashville, Tennessee, last week reached a preliminary settlement of \$10 million in a class action lawsuit brought by pregnant women who were given radioactive tracers fifty years ago and their relatives.

More than 800 women took part in studies of iron-deficiency anaemia sponsored by the university and the Tennessee Department of Health between 1945 and 1949. The university will pay them and their surviving relatives \$9.1 million; the Rockefeller Foundation, which also funded the studies, will pay the remainder.

The plaintiffs claimed that they and their children suffered because of the experiments, in which women at the Vanderbilt Prenatal Clinic ingested radioactive iron. But Jeff Carr, Vanderbilt's general counsel, said the studies contributed to the scientific understanding and prevention of “a significant nutritional problem at the time”.

Questions raised over fenfluramine tests

[WASHINGTON] The chairman of a congressional committee has launched an investigation into how the National Institutes of Health (NIH) and the Food and Drug Administration approved non-therapeutic experiments with fenfluramine, a now-banned diet drug, in children in New York (see *Nature* 392, 747 & 853; 1998).

In a letter to Harold Varmus, the head of the NIH, Dan Burton (Republican, Indiana), chairman of the House Government Reform and Oversight Committee, says that, in general, normal, healthy children are not supposed to be subjected to non-therapeutic, potentially dangerous experimental research. He also wants to know why research showing that therapeutic doses of fenfluramine can cause severe and sometimes permanent impairment in serotonin nerve cells in animals “[was] ignored, or how was it excused?”

Fusion reactor redesign ‘could halve costs’

[TOKYO] A new set of design parameters for the planned International Thermonuclear Experimental Reactor (ITER) has been proposed by a working group set up in February to investigate ways of cutting costs by relaxing some of the technical objectives (see *Nature* 391, 833; 1998).

At a meeting last week in Naka, Japan, the working group agreed that the estimated

US\$10 billion cost could be cut by up to half by reducing the radius of ITER's doughnut-shaped confinement vessel from 8 metres to 6 metres and by making various other changes. Although the new design parameters may no longer be sufficient to achieve ignition, the working group said it was confident that they would still allow many of the reactor's main technical goals to be met. The proposals will be discussed at the ITER council in Tokyo this month.

Emperor honoured by Britain's Royal Society

[LONDON] Emperor Akihito of Japan, an active scientist who has published extensively on the Japanese gobi fish and been closely associated with the support of science in Japan, was last week awarded the Royal Society's first King Charles II medal during a state visit to Britain.

The medal has been instigated by the society to give recognition to foreign heads of state “who have made an exceptional contribution to the promotion of science and its place in society”. Emperor Akihito has been closely involved with meetings of the Japan Academy, as well as major international awards such as the Japan prize and the Kyoto prize.

Small companion raises some big questions



[WASHINGTON] Planet, brown dwarf, or something else? Whatever TMR-1C is, astronomers have not seen anything like it before. Susan Terebey of the Extrasolar Research Corporation in Pasadena, California, was using the Hubble Space Telescope to study star formation in Taurus when she turned up this near-infrared image of a binary protostar with a jet of material pointing towards a small companion object (lower left). Terebey and her colleagues believe the object was gravitationally ejected from the system, and is most likely a newly formed planet about two to three times the mass of Jupiter. Ground-based spectroscopy should help settle the question of its origin.

S. TEREBEY/NASA

How big breweries failed the taste test

Sir — Two points mentioned in a recent book review on beer¹ deserve clarification. The term 'real ale' does need to be distinguished from other beers. Real ale refers specifically to an ale made from traditional ingredients in which carbonation is produced by means of secondary fermentation in the cask (or container) from which it is served².

A real ale uses no external source of carbon dioxide pressure to drive beer from the cellar to the tap, but instead employs a traditional hand-operated pump called a beer engine. The result is a beer with a much lower level of carbonation than that usually encountered in other beers, which might well be loosely defined as alcoholic beverages made with water, malted barley and hops, and fermented by yeast. A real ale requires the careful management by skilled publicans of a live beer whose flavour may change subtly with each passing hour or few pints served.

The preservation and revival of this method of preparing and serving beer has been largely due to the Campaign for Real Ale (CAMRA) in England, formed in 1971.

As mentioned in the book review¹, these activities have indeed led in part to the growth of this art in England, but have also played a decisive role in the evolution of the craft (pub, small/micro- and home-) brewing industry throughout the world.

Second, the brewing of mass-produced beers probably should not have been referred to as "sophisticated". Historically the largest breweries in the world were easily able to out-compete smaller breweries and public houses from the end of the prohibition years until the beginnings of the craft brewery movement by producing beer more inexpensively than smaller breweries (by virtue of the large volumes produced), and through extensive advertising campaigns. Along the way, though, the drive to increase the profit margin of beer and reduce the costs of its production led to the excessive use of brewing adjuncts (cheaper sources of fermentable sugars for yeast fermentation than malted barley) and to excessive dilution of the beers produced by means of high-gravity brewing techniques (reviewed in ref. 3).

This increased "sophistication" in the brewing industry led inevitably to consumer demand for better beers (if not indeed for 'real beer'). This worldwide trend in which beer consumers increasingly demand a quality product has led to significant losses in sales by the largest breweries, which are now making efforts to rectify this by introducing all-malt and other speciality beers to compete in this growing consumer sector. So, if anything, the "sophistication" of the large breweries has been responsible for their own misfortune.

Victor E. Buckwold
Ricardo Amils

*Centro de Biología Molecular,
"Severo Ochoa" Universidad
Autónoma de Madrid,
Cantoblanco,
28049 Madrid, Spain
e-mail: vebuckwold@trasto.cbm.uam.es*

1. Postgate, J. *Nature* **393**, 129 (1998).
2. Daniels, R. *Designing Great Beers: The Ultimate Guide to Brewing Classic Beer Styles* (Brewers Publications, Boulder, CO, 1996).
3. Buckwold, V. E. *Brewing Techniques* (in the press).

Bullying of PhDs

Sir — The 'feudal' master-servant relationship existing between a PhD supervisor and his or her student¹ has another facet seldom broached by academics. That is bullying. Employment legislation prohibits bullying at work, but, because PhDs are not salaried or contracted, they are not legally 'employees' and so are vulnerable to capricious supervisors.

I regret to say that the conduct of my PhD supervisor was tantamount to bullying. Corroborative complaints by peers and by me proved futile, culminating in my supervisor misappropriating corresponding authorship after editorial review of our manuscript. Although nebulous commitments to PhD supervision published in guidelines are welcome, they are merely cosmetic unless enforced impartially against the occasional aberrant supervisor. Experience has left me disaffected with my university, which is ostensibly content to allow a rogue supervisor to usurp authorship and confidence by allowing vulnerable PhDs to be bullied.

PhDs may now, however, be able to seek alternative recourse. In an unprecedented move, the British High Court has granted a student at the University of Cambridge, Mr Beg, judicial

review to challenge internal academic procedures². Mr Beg was allegedly denied an MPhil because he criticized a professor. If future legal challenges are to be avoided, reform is required to make possible equitable adjudication in alleged cases of supervisor misconduct. If UK institutions insist on maintaining the status quo, the courts may now intervene and universities will increasingly become embroiled in unwelcome litigation³. Denying the existence of bullying could become costly. Universities competing for funding and kudos can ill-afford to risk harbouring known aggressors, thus condoning their conduct and bringing departments into disrepute.

Inevitably, 'whistleblowers' (whether on matters of personal or academic misconduct) risk damage to their careers⁴. To ensure scientific integrity, postgraduate students need adequate protection from the repercussions of 'speaking out'. PhDs should surely be protected from bullying and unfair termination of studentships, in the same way as 'employed' researchers are protected by legislation and contracts.

Name and address supplied

1. McComb, J. *Nature* **387**, 448 (1997).
2. *The Guardian* 27 January 1998.
3. Robinson, J. *Education and the Law* **9** (2), 93–107 (1997).
4. Macilwain, C. *Nature* **385**, 669 (1997).
5. Poon, P. J. *Law, Med. Ethics* **23**, 88–95 (1995).

Summing up a wizard at maths ... warts and all

Sir — In his review of my book *Wizard: The Life and Times of Nikola Tesla* (*Nature* **388**, 135–136; 1997), L. Pearce Williams says that "no evidence is presented that [Tesla's] mathematical competence was at the level of contemporary mathematicians or even his fellow 'electricians'". I find this claim difficult to understand.

In 1937, Tesla was nominated for a Nobel prize in physics for his fundamental equations explaining AC polyphase systems, and his application of mathematical principles to his various inventions was acknowledged by Ambrose Fleming, Lord Kelvin, Ernest Rutherford, W. H. Eccles, Niels Bohr and Albert Einstein, among others.

Also, the statement that I "always" portray Tesla in a favourable light is incorrect. His financial deception of John Jacob Astor, his breach of contract with J. P. Morgan and his self-destructive tendencies and their link to the death of his older brother are just some of Tesla's quirks, mistakes, foibles or miscalculations that I pointed out.

Marc J. Seifer

*MetaScience Productions,
Box 32,
Kingston, Rhode Island 02881, USA*

It's a small world

James J. Collins and Carson C. Chow

The concept of Six Degrees of Separation has been formalized in so-called 'small-world networks'. The principles involved could be of use in settings as diverse as improving networks of cellular phones and understanding the spread of infections.

A few years ago, on American campuses, it was popular to play Six Degrees of Kevin Bacon. In this game, participants attempt to link the actor Kevin Bacon to any other actor through as few common films and co-stars as possible. Links are formed directly between Bacon and another actor if they appeared in the same film or indirectly through a chain of co-stars in different films (Fig. 1).

In the world of mathematics, a similar amusement involves assessing one's Erdős number, which measures the number of links needed to connect one to the prolific mathematician Paul Erdős through jointly authored papers. For example, individuals have an Erdős number of 1 if they co-authored a paper with Erdős. If one of their co-authors wrote a paper with Erdős, then they have an Erdős number of 2, and so forth. It has been pointed out¹ that Dan Kleitman has a combined Erdős/Bacon number of 3 because he wrote a paper with Erdős and appeared in *Good Will Hunting* with Minnie Driver, who appeared with Bacon in *Sleepers*.

These games are related to the popular concept of Six Degrees of Separation², which is based on the notion that everyone in the world is connected to everyone else through a chain of at most six mutual acquaintances. If two people have one mutual acquaintance, then they have one degree of separation. The estimate of six degrees of separation, which is related to the small-world phenomenon^{3,4}, arises from pioneering empirical work by Milgram³ and can be understood heuristically from a somewhat unrealistic assumption of random connectivity. That is, if each person knows about one hundred individuals, and given that there are about a billion people on the Earth, then seven connections or six degrees of separation are enough to link everyone together.

On page 440 of this issue⁵, Watts and Strogatz formalize this idea in what they call small-world networks. They demonstrate through numerical simulations that a network need not be very random to get this small-world effect. They consider a connected network with nodes and links. In the friendship analogy, each node represents a person and each link represents a single connection to an acquaintance. They then define

two measures. The first is a characteristic path length. This is the smallest number of links it takes to connect one node to another, averaged over all pairs of nodes in the network. The second measure is the clustering coefficient. This measures the amount of cliquishness of the network, that is, the fraction of neighbouring nodes that are also connected to one another. For example, in an all-to-all connected network, the clustering coefficient is one.

An example of a large-world network is one that is regularly and locally connected like a crystalline lattice. Such a network is highly clustered and the characteristic path length is large, scaling with the typical linear dimension of the network. On the other hand, a completely random network is poorly clustered and the characteristic path

length is short, scaling logarithmically with the size of the network.

What Watts and Strogatz⁵ do is to shift gradually from a regular network to a random network by increasing the probability of making random connections from 0 to 1 (see Fig. 1, page 441). They then measure the characteristic path length and the amount of clustering of the network as a function of the amount of randomness. They find that path length and clustering depend differently on the amount of randomness in the network. The characteristic path length drops quickly, whereas the amount of clustering drops rather slowly. This leads to a small-world network in which the amount of clustering is high and the characteristic path length is short. So a small world can exist even when the cliquishness is imperceptibly different from that of a large world.

The explanation for this effect is that it only takes a few short cuts between cliques to turn a large world into a small world. In the friendship analogy, it only takes a small number of well-connected people to make a world small. The interesting and surprising thing is that it is impossible to determine whether or not you live in a small world or a large world from local information alone. The average person (node) is not directly associated with the key people (the clique-linkers).

Small-world connectivity has consequences that could be good or bad,



Figure 1 Three degrees. Because Kevin Bacon has appeared in many films, most actors have low Bacon numbers and the game Six Degrees of Kevin Bacon has declined in popularity. It is possible to centre the game around a newer star such as Leonardo DiCaprio. These film stills, running clockwise, show that in this case there are at most three degrees of separation between DiCaprio and Helena Bonham-Carter, through Kate Winslet (*Titanic*, Columbia TriStar; *Sense and Sensibility*, Columbia TriStar), Emma Thompson (*Sense and Sensibility*; *Much Ado About Nothing*, Entertainment Films) and Kenneth Branagh (*Much Ado About Nothing*; *Frankenstein*; Columbia TriStar). Short cuts between cliques could be created in this game through some of DiCaprio's well-connected co-stars such as Sharon Stone (*The Quick and the Dead*; TriStar; not shown).

AP/THE KOBAL COLLECTION

depending on the system and circumstances. In a simple model of disease spreading, for example, Watts and Strogatz show that the time for global infection behaves much like the characteristic path length. So it takes only a few short cuts to increase the spreading of disease significantly. Clearly, in this case, small-world coupling is problematic.

The findings of Watts and Strogatz could be put to good use, however, particularly in the case of existing networks. For instance, it may be possible to improve the performance of cellular-phone networks by deliberately introducing a few random connections between cells (nodes). Such a change could improve traffic flow around the network, without requiring the creation of an entirely new set of relay stations. Similar dynamics could also be exploited to improve the flow of information throughout the Internet. Although we are not in a position to redesign the Internet from scratch, it is possi-

ble to introduce a few random links between nodes along the backbone of the Internet. These small-world modifications could substantially reduce the time needed to send a message by electronic mail or find a particular Web site, as well as improve the reliability of the overall network. Strategies for determining and achieving optimal small-world connectivity remain to be developed. □

James J. Collins and Carson C. Chow are in the Center for BioDynamics, Department of Biomedical Engineering and Department of Mathematics, Boston University, 44 Cummington Street, Boston, Massachusetts 02215, USA.

e-mail: collins@enga.bu.edu

1. <http://www.acs.oakland.edu/~grossman/Erdoshp.html>
2. Guare, J. *Six Degrees of Separation: A Play* (Vintage, New York, 1990).
3. Milgram, S. *Psychol. Today* **2**, 60–67 (1967).
4. Kochen, M. (ed.) *The Small World* (Ablex, Norwood, NJ, 1989).
5. Watts, D. J. & Strogatz, S. H. *Nature* **393**, 440–442 (1998).

Marine ecology

Microbial microdiversity

Jed A. Fuhrman and Lisa Campbell

The bacterial group *Prochlorococcus*, discovered only a decade ago, may be the most abundant component of phytoplankton in the sea. These tiny (0.6 µm) organisms uniquely contain the photosynthetic pigments divinyl chlorophyll *a* and *b*, and are major primary producers in tropical and subtropical waters (that is, some 75% of the world's oceans). They contribute between 10% and 80% of total local primary production^{1,2}.

One apparent reason for their success is their ability to grow well over a wide range of light conditions. So, how do they do it? As reported by Moore *et al.* on page 464 of this issue³, it turns out that the feat is accomplished by the coexistence of closely related populations that are genetically adapted to remarkably different light levels. Close relatives from the same sample can be so different in their light response that one may grow optimally in light that is bright enough to inhibit the other completely. Together with the results of Urbach *et al.*⁴, these observations not only explain why *Prochlorococcus* is so successful, but they also bear on issues that are central to the measurement and interpretation of microbial diversity.

To understand these broader implications, one must first realize how such studies have recently changed. Classical identification of a microbe requires cultivation of the organism concerned, which is difficult or impossible for most organisms from nature (perhaps 99%). In consequence, although as many as 5,000 bacterial and archaeal species are officially recognized, the true number is probably in the millions⁵. There is simply no

practical way to carry out a comprehensive survey of natural microbial diversity by classical techniques alone.

Norman Pace and colleagues set out to find a way around this limitation. They showed⁶ that one could clone the genes for 16S ribosomal RNA from a naturally occurring, mixed, microbial biomass, sequence them, and then identify the organisms from which the genes came by comparing the sequences to the large database of 16S rRNA sequences. Phylogenetic analysis places the clone sequences on a tree, from which relationships to known organisms or other environmental clones can be seen. By eliminating the need for cultivation, this approach has been a remarkably fruitful way to survey diversity. It has led to the discovery of major groups of organisms, still uncultured and without formal names, such as the low-temperature archaea^{7–9}. In the past year alone, more than 20 new divisions of bacteria have been reported^{9–11} at the phylum — or even possibly the kingdom — level.

The cloning approach has also yielded several clusters of closely related sequences, which may indicate that diversity is common on much smaller scales. Examples of marine clusters include SAR 11 (ref. 12) and Marine Group I archaea^{7–9} (Fig. 1), both of which are distributed widely around the world yet are evolutionarily distant from all known cultures. Typical within-cluster differences in 16S rRNA sequences are <1–10%, and clades within clusters are also apparent^{9,12}. But it is difficult to know what the within-cluster diversity means without having living cultures to examine. Are relatively small differ-

ences in sequences (termed microdiversity by Moore *et al.*³) the result of genetic drift or microevolution in physiologically similar clonal populations, or do they represent macroevolution of distinct populations with genetically determined adaptations and different niches? The answer to this question is crucial, because it tells us whether or not it is appropriate to lump together the members of the cluster for the purposes of diversity analysis. The difference could multiply our perception of natural diversity by several times.

Moore and colleagues' study³ is exciting because it concentrates on a cluster that has been detected repeatedly in open ocean environments by the cloning approach¹³, but which can also be studied in nature by flow cytometry or in laboratory culture. This group has been called the Marine Picophytoplankton Clade, and includes *Prochlorococcus* and *Synechococcus* cyanobacteria⁴. These organisms do not have the same light-harvesting components (*Synechococcus* possess true chlorophyll *a* and phycobilisomes, and lack divinyl chlorophyll *a* and *b*), but they are about 96–98% identical in their 16S rRNA sequences. Phylogenetic analysis shows identifiable sub-lineages within this group, with one being the clade of *Prochlorococcus* that is adapted to high light intensities^{3,4}. Members of this clade have about 97–98% 16S rRNA sequence identity with *Prochlorococcus* adapted to low light levels, but the two types coexist³.

Moore *et al.*, then, show how differences of only about 2% in 16S rRNA sequence correspond to ecologically significant physiological diversity. So if lumping together clusters with a 2% difference would be a mistake, what about even smaller differences? Examination of the 16S rRNA database from cultures shows many examples where there are notable physiological differences despite tiny (<0.5%) sequence differences (for example, the anthrax organism *Bacillus anthracis* and the insect pathogen *B. thuringiensis*¹⁴); and with plasmids or other mobile genetic elements often coding for important properties such as virulence, there are instances where organisms with even completely identical 16S rRNA sequences occupy different niches¹⁴.

Estimates of microbial diversity based on 16S rRNA probably therefore represent a minimum of the true ecological situation. There are other factors to be considered, however. For instance, a small part of the apparent diversity in cloned clusters is possibly due to variations in rRNA-coding gene sequences between the multiple copies typically present in many organisms. Or a small part may be due to experimental errors in carrying out the polymerase chain reaction, and cloning and sequencing¹². Although these points are not new, they need to be

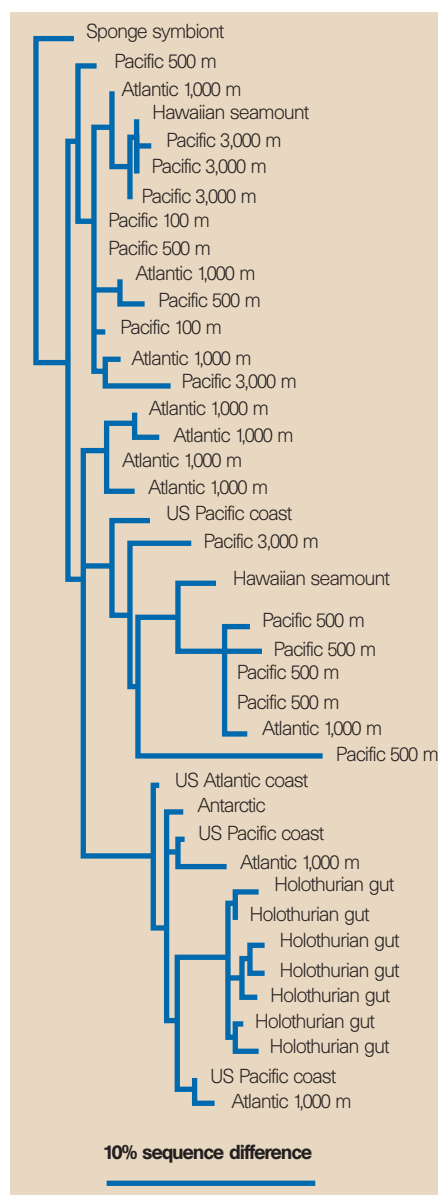


Figure 1 An example of microbial diversity for which the ecological and physiological correlates are not known. These are representatives of the Marine Group I Archaeal Cluster, and they occur in widespread locations and at different depths. The 16S ribosomal RNA sequences used to create this phylogenetic tree came from clones obtained directly from field samples; no cultures are known. The nearest cultured relatives, with only about 70% sequence similarity to this cluster, are extreme thermophiles, but these clones do not represent thermophiles. The meaning of this small-scale diversity, which probably represents adaptive radiation of species, can at the moment only be inferred from studies such as that by Moore *et al.*³ discussed here. (Figure modified from ref. 9. The scale applies only to the horizontal relationships.)

kept in mind. For example, screening with restriction-fragment length polymorphisms, which is sometimes used to avoid 'duplicates' in 16S rRNA cloning studies, may remove functionally different types of

organism from consideration because it usually does not distinguish between small differences in sequences.

The cloned 16S rRNA sequences provide a new window into microbial diversity and the species concept for microbes. We can think of cloned sequence clusters as reflecting adaptive radiations of species, analogous to those in multicellular organisms^{12,15}. But 16S rRNA sequences alone are insufficient to delineate species and are better for broader groups¹⁴, so cloning surveys are only a first step. Ward¹⁵ suggests a revision of the conceptual framework for defining microbial species to use more natural (evolutionary and ecological) criteria, as is the case with multicellular organisms.

In his classic 'Paradox of the Plankton', put forward decades ago, Hutchinson asked how it is possible to maintain so many phytoplankton species in a homogeneous environment with only a few potentially limiting resources¹⁶. The newly recognized bacterial diversity makes matters even more complicated. Although there have been many possible explanations, including the suggestion that ubiquitous viruses may help to maintain diversity¹⁷, the paradox still perplexes us today. It also reminds us how much we have yet to learn about biodiversity.

Jed A. Fuhrman is in the Department of Biological Sciences and the Wrigley Institute for

Environmental Studies, University of Southern California, Los Angeles, California 90089-0371, USA.

e-mail: fuhrman@usc.edu

Lisa Campbell is in the Department of Oceanography, Texas A&M University, College Station, Texas 77843-3146, USA.

e-mail: lcampbell@ocean.tamu.edu

- Goerick, R. & Welschmeyer, N. A. *Deep-Sea Res.* **40**, 2283–2294 (1993).
- Liu, H., Nolla, H. A. & Campbell, L. *Aquat. Microb. Ecol.* **1**, 39–47 (1996).
- Moore, L. R., Rocap, G. & Chisholm, S. W. *Nature* **393**, 464–467 (1998).
- Urbach, E., Scanlan, D. J., Distel, D. L., Waterbury, J. B. & Chisholm, S. W. *J. Mol. Evol.* **46**, 188–201 (1998).
- Hawthornthwaite, D. L. & Kalin-Arroyo, M. T. in *Global Biodiversity Assessment* (eds Heywood, V. H. & Watson, R. T.) 107–192 (Cambridge Univ. Press, 1995).
- Pace, N. R., Stahl, D. A., Lane, D. J. & Olsen, G. J. *Adv. Microb. Ecol.* **9**, 1–55 (1986).
- Fuhrman, J. A., McCallum, K. & Davis, A. A. *Nature* **356**, 148–149 (1992).
- DeLong, E. F. *Science* **280**, 542–543 (1998).
- Fuhrman, J. A. & Davis, A. A. *Mar. Ecol. Prog. Ser.* **150**, 275–285 (1997).
- Borneman, J. & Triplett, E. W. *Appl. Environ. Microbiol.* **63**, 2647–2653 (1997).
- Hugenholtz, P., Pittule, C., Hershberger, K. L. & Pace, N. R. *J. Bacteriol.* **180**, 366–376 (1998).
- Field, K. G. *et al.* *Appl. Environ. Microbiol.* **63**, 63–70 (1997).
- Mullins, T. D., Britschgi, T. B., Krest, R. L. & Giovannoni, S. J. *Limnol. Oceanogr.* **40**, 148–158 (1995).
- Paly, T., Nakamura, L. K. & Cohan, F. M. *Int. J. System. Bacteriol.* **47**, 1145–1156 (1997).
- Ward, D. M. *Curr. Opin. Microbiol.* **1**, 271–277 (1998).
- Hutchinson, G. E. *Am. Nat.* **45**, 137–145 (1961).
- Fuhrman, J. A. & Suttle, C. A. *Oceanography* **6**, 51–63 (1993).

Mathematics

The big match

Ivar Ekeland

One of the oldest problems in game theory, known as the big match, has just been solved by the French mathematician Nicolas Vieille. Let me tell it as a story.

Once upon a time there was a king, powerful and wise. He called in his trusted minister and told him: "Today I will leave, and in my absence, I put you in charge of my kingdom. You will not hear from me again until the day of my return, if I ever come back. On that day, if I find you hard at work at my desk, you will become king in my place, and you will live in this palace for ever and ever. If, on the other hand, I find you idling, enjoying the perks of office, you will be put in chains and imprisoned in a dungeon where you will be tormented for ever and ever. Be aware that every night I will be informed of how you spent the past day, and I can always come back the next day."

What should the minister do, assuming that the king will use the system to his own advantage? If he spends the whole time working, then the king, being informed that he is at his desk every day, will expect him to go on showing up every morning for work, and will not come back, so the minister can only look

forward to a lifetime of work. If he spends every day enjoying the good life, he can be assured that the king, being duly informed of his misdemeanour, will come back to chastise him and will indeed find him idling, so that he can look forward to the bottomless pit. In neither case will he have any chance of catching the king out by being at work on his return, and so become king himself.

Somewhere in between the two extremes of 'always work' and 'always shirk', there must be a compromise that blends working and shirking in an optimal manner. But this compromise is hard to find. Every day, as he wakes up, the minister has to balance the pleasure of shirking against the risk of the king coming back today, based on his information yesterday. He must also bear in mind that by shirking today, he is increasing the likelihood that the king will come back tomorrow. As he wakes up tomorrow, he may judge the new risk unacceptable, and find himself forced to go to work; in other words, by shirking today, he may be committing himself to work tomorrow.

His difficulties, as always in game theory, are compounded by the fact that the king is perfectly able to put himself in his minister's



shoes. So if, for instance, there were a way for the minister to reach a definite conclusion as to whether he should work or shirk today, the king, having precisely the same information (namely, what the minister did yesterday and the days before), would reach the same conclusion, and take advantage of it — returning if the minister planned to shirk, and staying away if he planned to work. The way out is to resort to probabilistic strategies: as he wakes up, the minister decides on certain probabilities with which to work or shirk, and then casts lots to find out what he will actually do. Similarly, the king will throw dice every morning to find out whether to return or to stay away, only needing to decide the odds for returning or staying away.

This problem was first stated by Gillette¹ in 1957. Its solution depends on the values the minister and the king attach to the different outcomes, and also on the way they value the future, compared with the present. In 1968, Blackwell and Ferguson² found the solution to the particular case when the king and the minister have exactly opposite interests (say, for instance, that if the minister is caught, he can stay out of jail by paying every day a large amount of money M to the king, and that if the king is caught out, he can keep his kingdom by paying the minister this same amount M every day). They proved that, if everyone holds a long-range view, the best strategy for the king is to choose at the outset some large number N , and then to compute every morning the number $D = N + A - B$, where A is the number of days the minister has worked since the game began and B the number of days he has shirked. The king then assigns a probability $p = 1/D^2$ to coming

back and throws dice to that effect. As for the minister, his best policy is to flip a fair coin every morning and to work if it comes up tails. Note that, as long as A is close to B , the probability of the king returning today is close to the constant $1/N^2$.

After this solution, things stalled for ten more years, until Mertens and Neyman^{3,4} extended the result to games with more than two possible outcomes, and where chance comes in more complicated ways. But they

were still restricted to zero-sum games, that is, where one player's loss is exactly the other's gain. Non-zero-sum games have more general payoffs, allowing both players to win in some outcomes, and thereby encouraging some degree of cooperation. But the question of whether such games also could be solved eluded researchers for 20 more years.

Finally, Vieille has proved that they can all be solved. It is an extremely long proof, which spans a series of four difficult papers^{5–8}. It is not a constructive proof, in the sense that in most cases one would not be able to actually give the solution, as I did in the preceding story: one would just know that it exists. However, it is an outstanding achievement, requiring extremely sophisticated tools from all sides of mathematics, from ergodic theory to algebraic geometry. It gives us a new class of models with which to study conflict and cooperation — models that will certainly find their way into the study of social interactions and economics. □

Ivar Ekeland is at the University of Paris-Dauphine, 75775 Paris, Cedex 16, France.

e-mail: ivar.ekeland@dauphine.fr

1. Gillette, D. *Ann. Math. Stud.* **39**, 179–187 (1957).
2. Blackwell, D. & Ferguson, T. S. *Ann. Math. Stat.* **39**, 159–163 (1968).
3. Mertens, J. F. & Neyman, A. *Int. J. Game Theory* **10**, 53–66 (1981).
4. Mertens, J. F. & Neyman, A. *Proc. Natl Acad. Sci. USA* **79**, 2145–2146 (1982).
5. Vieille, N. *Int. J. Game Theory* **21**, 395–404 (1993).
6. Vieille, N. *2-Players Stochastic Games 1: A Reduction*, preprint, Univ. Paris-Dauphine (1997).
7. Vieille, N. *2-Players Stochastic Games 2: The Case of Recursive Games*, preprint, Univ. Paris-Dauphine (1997).
8. Vieille, N. *Large Deviations and Stochastic Games*, preprint, Univ. Paris-Dauphine (1997).

Immunology

Licence to kill

Antonio Lanzavecchia

When the body's own cells become infected, the immune system can take the extreme step of killing them to preserve the rest of the organism. It is no surprise that this potentially dangerous mechanism has to be kept under strict control, and on pages 474¹, 478² and 480³ of this issue, three groups suggest a new mechanism by which this control operates.

The immune system's T-killer cells exist as inactive precursors that must be activated to develop killing capacity. Activation entails recognition of antigenic peptides bound to class I molecules of the major histocompatibility complex (MHC) on the surface of antigen-presenting cells (APCs). But this precursor–APC interaction is often not enough to stimulate the T-killer cell (Fig. 1a, overleaf) and, in these cases, a T-helper cell that can recognize an antigen on the same APC is also required⁴.

The need for helpers and the nature of this help have intrigued immunologists for years. The first model (Fig. 1b) suggested that T-helper and T-killer cells recognize their specific antigens simultaneously on the same APC, and that cytokines (such as interleukin-2) produced by the activated T-helper cell would act on the T killer and facilitate its response^{4,5}. Although interleukin-2 can substitute for helper cells in some systems, there are two main problems with this model. First, it requires that two rare antigen-specific cells meet on the same antigen-bearing APC — an event that has an extremely low probability. Second, and more importantly, some killer responses can be elicited in the absence of helper cells⁶, implying that the need for help is conditional rather than absolute.

The alternative theory, proposed by Polly Matzinger on theoretical grounds, is a 'licensing' model. Matzinger suggested that

by recognizing antigen on APCs, T-helper cells deliver a signal that activates the APCs⁷. The licensed APC can then directly stimulate T-killer cells (Fig. 1c). The key to this model is a specialized 'professional' APC, which has been identified as the dendritic cell. Dendritic cells sample antigens in all body tissues and migrate to lymph nodes where they present antigens to T cells⁸. The molecules responsible for the interaction between the T-helper and dendritic cells are called CD40L and CD40. CD40L is a membrane molecule expressed by antigen-stimulated T-helper cells. It interacts with and triggers CD40, a surface receptor that can activate dendritic cells^{9–11}.

The three new papers support this model of T-killer-cell activation. Bennett *et al.*² and Schoenberger *et al.*³ show that mice lacking T-helper cells cannot mount a killer response when they are injected with APCs that display an antigen recognized by T-killer cells. Remarkably, however, the authors could induce a killer response by injecting the mice with an antibody against CD40 — by binding CD40 and triggering the APCs, the anti-CD40 antibody acts as a surrogate of T-helper cells. Moreover, Schoenberger *et al.* noticed that, in normal mice, they could block the induction of a helper-dependent T-killer response using antibodies to CD40L. This inhibition was reversed by adding the stimulatory anti-CD40 antibody. Bennett and colleagues also showed that mice lacking either CD40 or CD40L cannot mount helper-dependent T-killer responses.

To define the stimuli that enable the APC to activate T-killer cells, Ridge *et al.*¹ manipulated dendritic cells *in vitro*. They show that dendritic cells carrying specific antigen cannot stimulate T-killer responses *in vitro* or *in vivo* unless they are first stimulated by anti-CD40 antibodies or specific T-helper cells. The interaction between T-helper cells and APCs can be dissociated in time from the interaction between APCs and T killers. So, the rare antigen-specific T cells do not need to interact with APCs simultaneously — they can do it sequentially.

These results^{1–3} clearly illustrate that the interaction between CD40L and CD40 can be crucial in bringing the APCs to a state in which they can autonomously trigger T-killer responses. Interestingly, however, they also suggest that there may be alternative ways to activate APCs. Bennett *et al.* show that a T-killer response to the same antigen may or may not depend on the T-helper cells, according to whether the antigen is administered on cells or in association with an adjuvant (a substance that induces inflammation). In addition, Ridge *et al.* show that dendritic cells can also be activated after infection with influenza virus.

What happens to dendritic cells during this activation, and why do different pathways lead to the same effect? It is well known that dendritic cells need to be activated to perform

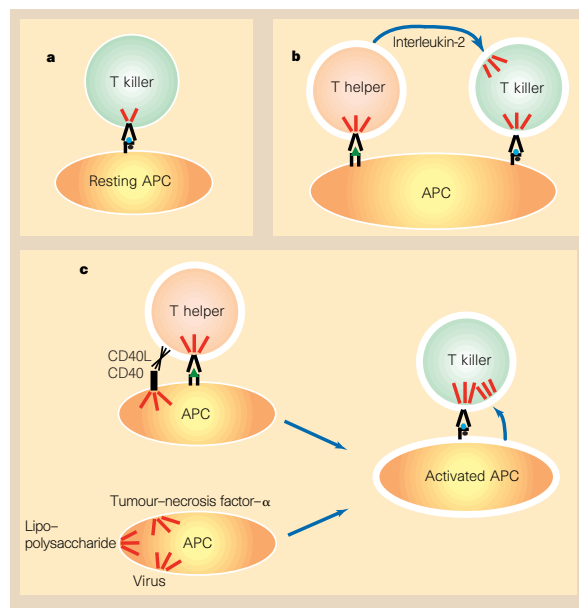


Figure 1 Antigen-presenting cells (APCs) need a licence to help T-killer cells. **a**, T-killer precursors recognize antigen on resting APCs, but do not receive enough signal to be activated. **b**, According to the traditional model, T helpers and T killers recognize antigen on the same APC. The helper is activated and produces interleukin-2 which contributes to activation of the killer. **c**, In the new model, supported by the experiments of Ridge *et al.*¹, Bennett *et al.*² and Schoenberger *et al.*³, the APCs are licensed to activate T-killer cells by T helpers or by other stimuli.

optimally. In their resting state, they possess low levels of MHC and co-stimulatory molecules, so they are poor at stimulating T-killer cells. Resting dendritic cells can be activated by inflammatory cytokines and bacterial products that upregulate MHC and the co-stimulatory molecules, thereby increasing the ability to stimulate T cells⁹. Furthermore, activation of dendritic cells by CD40 results in increased production of cytokines, especially interleukin-12 (refs 10, 11). The latest findings^{1–3} can thus be explained by the fact that co-stimulatory molecules and interleukin-12, which are well-known requirements for activation of T-killer cells, need to be induced in dendritic cells by T-helper cells and inflammatory stimuli.

The licensing model reconciles, in a simple concept, many apparently contrasting observations. For instance, it explains how the response to the same antigen can be T-helper dependent or independent. In the presence of adjuvant (which Charlie Janeway calls "the little dirty secret of immunologists"¹²), the inflammatory cytokines generated may be enough to activate dendritic cells to a state where they can autonomously stimulate a T-killer response, even in the absence of T-helper cells. On the other hand, if antigen is presented by resting dendritic cells, the T-helper-mediated CD40 triggering would be absolutely required. But this model is still compatible with the possibility that particularly strong antigens might be able to stimulate T-killer responses, even in the absence of help or inflammation.

The licensing concept has further implications for the generation of polarized immune responses. Activation of the APCs and production of interleukin-12 are required for the T-helper responses that protect from intracellular pathogens. Conversely, suppressive T cells might inhibit or modulate the function of the dendritic cells, and in this way interfere with the immune response.

Thus, rather than being chaperones that bring T helpers and T killers together, the dendritic cells have now become active protagonists that are licensed to activate killer cells. Stimuli from pathogens, inflammatory cytokines and T-helper cells are integrated in dendritic cells, which ultimately tune the capacity to stimulate T-cell responses. It makes sense to have different ways to activate dendritic cells so that an appropriate response to any offending agent can be mounted effectively.

The possibility of manipulating the activation of dendritic cells will open new avenues for vaccination and therapy. By stimulating dendritic cells, the response to weak antigens such as tumour-specific antigens (which are normally encountered outside an inflammatory context) may be boosted. This could be done using anti-CD40 antibodies, as proposed by Ridge *et al.*¹, Bennett *et al.*² and Schoenberger *et al.*³. Alternatively, as others have suggested¹¹, dendritic cells could be pulsed with an antigen that is recognized by the T-helper memory cells, thereby causing them to be activated in a timely fashion in lymphoid organs. □

Antonio Lanzavecchia is at the Basel Institute for Immunology, Grenzacherstrasse 487, CH 4005 Basel, Switzerland.
e-mail: lanzavecchia@bii.ch

1. Ridge, J. P., Di Rosa, F. & Matzinger, P. *Nature* **393**, 474–478 (1998).
2. Bennett, S. R. M. *et al.* *Nature* **393**, 478–480 (1998).
3. Schoenberger, S. P., Toes, R. E. M., van der Voort, E. I. H., Offringa, R. & Melief, C. J. M. *Nature* **393**, 480–483 (1998).
4. Keene, J. A. & Forman, J. J. *Exp. Med.* **155**, 768–782 (1982).
5. Mitchison, N. A. & O'Malley, C. *Eur. J. Immunol.* **17**, 1579–1583 (1987).
6. Buller, R. M., Holmes, K. L., Hugin, A., Frederickson, T. N. & Morse, H. C. *Nature* **328**, 77–79 (1987).
7. Guerder, S. & Matzinger, P. *J. Exp. Med.* **176**, 553–564 (1992).
8. Banchereau, J. & Steinman, R. M. *Nature* **392**, 245–252 (1998).
9. Sallusto, F. & Lanzavecchia, A. *J. Exp. Med.* **179**, 1109–1118 (1994).
10. Caux, C. *et al.* *J. Exp. Med.* **180**, 1263–1272 (1994).
11. Cella, M. *et al.* *J. Exp. Med.* **184**, 747–752 (1996).
12. Janeway, C. A. Jr *Cold Spring Harb. Symp. Quant. Biol.* **54**, 1–13 (1989).

Volcanoes

How did the summer go?

David M. Pyle

The summer of 1601 was, by all accounts, truly awful. Across England, "the month of June was very colde, frosts every morning"¹. It was a similar story in northern Italy, where freezing weather extended into July, and the sky was 'overcast' for much of the year². Iceland and Scandinavia also apparently experienced unseasonal weather. Just another poor northern European summer? It seems not.

Papers by Briffa *et al.*³ and de Silva and Zielinski⁴ (pages 450 and 455 of this issue) now reveal the true scale of this cold summer and the reasons for it. Patterns of tree growth across the Northern Hemisphere, analysed by Briffa *et al.*, confirm that the summer of 1601 was by far the coldest of the past 600 years, and about 0.8 °C cooler than the summer mean for the period 1881–1960. The cause is identified as a spectacular eruption of the volcano Huaynaputina in Peru in February 1600. Fieldwork^{4,5} now shows that Huaynaputina (Fig. 1) was one of the largest eruptions of the past few centuries. In their paper, de Silva and Zielinski calculate that it threw out nearly 10 km³ of molten rock, and left prominent records of sulphate pollution in polar ice-cores. For comparison, the damaging eruption of Mount Pinatubo, Philippines, in 1991, involved about 5 km³ of magma, whereas that of El Chichón, Mexico, in 1982 involved about 1.3 km³ of magma.

Eruptions of the scale of Huaynaputina

occur only once or twice a century, and are often followed by a widespread, but short-lived, climate response caused by the volcanic pollution. Volcanic sulphur dioxide injected into the stratosphere oxidizes to form a sulphate aerosol layer that coarsens slowly, and drifts around the globe. This aerosol layer intercepts incoming solar radiation, disturbing Earth's surface temperatures. The mid-latitudes of the Northern Hemisphere seem to be particularly sensitive to the effects of large eruptions⁶, experiencing slight winter warming and marked summer cooling as a result. Eventually, after transport to the poles, the sulphate aerosol may be preserved in accumulations of snow.

In detail, the time taken for sulphate to disperse from the tropics to higher latitudes depends on the phase of the stratospheric 'quasi-biennial oscillation'⁷. Under some conditions, aerosols can be stored within a tropical stratospheric reservoir for a year or two before being transported to higher latitudes. This introduces a time lag between large eruptions in the tropics and their effects at mid- to high latitudes. It also reduces the chance of identifying the source of any individual sulphate peak in polar ice-cores — which can only be done by matching the compositions of the accompanying fragments of volcanic glass with a known eruption^{4,8}. If volcanic ash is held above the tropics for long enough, the ash particles will



100 YEARS AGO

One of the elephants in Barnum and Bailey's Show, which has been visiting Liverpool during the past two weeks, having recently shown signs of insubordination, Mr. Bailey determined, in order to perfectly safeguard his visitors, to sacrifice the animal. He has had during his life occasion to destroy many elephants, which, as a rule, he has handed over to experienced veterinary and other surgeons, who have tried various methods, such as poisoning, shooting and bleeding. All have proved, however, unsatisfactory, because uncertain, tedious and not seldom dangerous to those engaged in conducting the operations. On this occasion it was determined, after consultation with several experts and with the Secretary of the Royal Society for the Prevention of Cruelty to Animals, to kill the elephant by strangulation, which had once before been adopted with success by Mr. Bailey. Accordingly it was arranged that on a recent Sunday morning — the day most suitable to the Show people and that freest from intrusion by the public — Don, as the doomed elephant, who was supposed to be about twenty-two years of age and nearly 4 1/2 tons in weight, was named, should be strangled. From *Nature* 2 June 1898.

50 YEARS AGO

A valuable means of bringing about active co-operation between industry and the universities has recently been shown by the United Steel Companies, Ltd. At the invitation of the chairman, Sir Walter Benton Jones, fourteen mechanical and production engineering professors from British universities spent four days ... in visiting the branches of the United Steel Companies at Sheffield, Scunthorpe and Stocksbridge, and in joint discussion both at the various works and at the Sheffield headquarters. ... [T]he research staff of the United Steel Companies were grateful for suggestions thrown out by the professors, while the latter were appreciative of the new problems put to them and which, in some cases, could usefully be followed up by them and their research staffs. By arranging this tour the United Steel Companies have shown a way in which industry and scholarship can work together for their mutual help and the well-being of British economy. From *Nature* 5 June 1948.



Figure 1 Craters of Huaynaputina, in the Peruvian Andes. This was the site of one of the largest eruptions of the past few centuries, in February 1600. From their reconstruction of the extensive deposits of volcanic ash around the volcano, de Silva and Zielinski⁴ calculate that this eruption threw out nearly 10 km³ of molten rock. Briffa *et al.*³ have studied patterns of tree growth, which show that the following summer of 1601 was the coldest of the past 600 years across the Northern Hemisphere. Sulphate pollution in polar ice-cores from the same time suggests that the eruption of Huaynaputina was directly responsible for this widespread cooling.

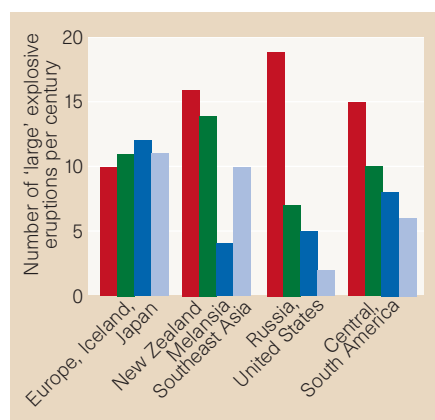


Figure 2 A compilation of the numbers of reported¹⁰ 'large' explosive volcanic eruptions — those with a volcanic explosivity index of 4 or more — during the past four centuries. (Red, twentieth; green, nineteenth; dark blue, eighteenth; light blue, seventeenth.) Whereas areas with long historical records (Europe, Iceland and Japan) experience eruptions at a constant rate, other areas apparently do not, showing that the record for these other regions is grossly incomplete. There were probably several tens of, as yet, unknown mid- to high-latitude eruptions in Kamchatka, the Kurile islands and Alaska between the seventeenth and nineteenth centuries. The seventeenth- and eighteenth-century records of tropical eruptions in Southeast Asia and Central and South America are also far from complete.

settle out rather than being transported to the poles.

In contrast, the products of mid-latitude eruptions are dispersed rapidly towards the pole. The climate response in mid-latitudes is more immediate, and polar ice-cores are more likely to preserve sulphate and ash together. A final consequence is that the effects of any two eruptions of similar size can be quite different, depending on when and where they occurred. All of these features can be recognized in the Northern Hemisphere tree-ring record, as described by Briffa *et al.*³

The 'year without a summer' in 1816 and the subsequent cold summers of 1817 to 1819 all feature in the 'top 30' most extreme Northern Hemisphere summers of the past six centuries. These followed the eruption of Tambora, Indonesia, in April 1815. Similarly, the cool summer of 1884 followed the August 1883 eruption of Krakatau, Indonesia. In contrast, the June 1912 eruption of Novarupta, Alaska, had an immediate cooling effect, and shows up in the tree-ring records from 1912 as the most extreme summer of the twentieth century so far. An eruption of comparable scale^{8,9}, at Santa Maria, Guatemala, in October 1902, failed to register as a significant disturbance in the tree-ring record.

The record of the violent, but non-explosive, eruption of Laki, Iceland, in 1783 is also

instructive. This eruption released more sulphur than Tambora, leading to a thick 'dry fog' that spread across northern Europe. However, because most of the pollution was restricted to the troposphere, cooling effects were locally severe but not hemisphere-wide, and the summer of 1783 ranks only twenty-sixth in the record of extreme low tree-ring density.

The association of the two most extreme summers, 1601 and 1816, with known eruptions suggests that some of the other cold summers might also have volcanic causes. Indeed, many of these, notably 1641–43, 1695 and 1698, match periods with increased levels of sulphate in polar ice-cores⁸. This is the 'smoking gun' that implicates volcanic eruptions as the cooling agent.

But which eruption? Tens of volcanoes erupt explosively every year. Most of them are too small to have any climatic effect, even though they might be locally devastating, and thereby spawn a rich historical written record. Unfortunately, the catalogue of known eruptions¹⁰ rapidly becomes incomplete further back in time. A glance at the known 'large' eruptions of the past 400 years (Fig. 2) shows that, because eruptions in areas with long historical records occur at a more or less constant rate, there must be many tens of 'large', and potentially globally

significant, eruptions that remain undiscovered from the remoter areas of the Pacific rim.

It took the combination of several different, highly time-resolved approaches (tree-ring and ice-core records) and field study to unravel the story of the summer of 1601. The largest volcanic eruptions need not be responsible for the largest sulphate releases⁹, however, and none of the direct or indirect indicators of past volcanism is complete, so reading the record of past eruptions and their climatic effects poses a continuing challenge. □

David M. Pyle is in the Department of Earth Sciences, University of Cambridge, Downing Street, Cambridge CB2 3EQ, UK.

e-mail: dmp11@cam.ac.uk

1. Stow, J. *The Annals of England Increased and Continued Until This Present Year 1605* (London, 1605).
2. Pavese, M. P., Banzon, V., Colacino, M., Gregori, G. P. & Pasqua, M. in *Climate since A.D. 1500* (eds Bradley, R. S. & Jones, P. D.) 155–170 (Routledge, London, 1992).
3. Briffa, K. R., Jones, P. D., Schweingruber, F. H. & Osborn, T. J. *Nature* **393**, 450–455 (1998).
4. de Silva, S. & Zielinski, G. A. *Nature* **393**, 455–458 (1998).
5. Thouret, J.-C. *et al.* *C. R. Acad. Sci.* **325**, 931–938 (1997).
6. Robock, A. & Mao, J. *J. Clim.* **8**, 1086–1103 (1995).
7. Trepte, C. R. & Hitchman, M. H. *Nature* **355**, 626–628 (1992).
8. Zielinski, G. A. *J. Geophys. Res.* **100**, 20937–20955 (1995).
9. Rampino, M. R. & Self, S. *Nature* **310**, 677–679 (1984).
10. Simkin, T. & Siebert, L. *Volcanoes of the World* (Geoscience Press, Tucson, AZ, 1994).

Neurobiology

A new image for fear and emotion

Steven E. Hyman

The study of emotion is enjoying a renaissance. Despite the importance of emotion in behaviour — indeed, in survival itself — it had largely been ignored as a field of study until recently. Then, several groups identified the anatomy, physiology and chemistry of the circuits that underlie fear in animals and humans, creating new possibilities for understanding emotions and human mental illness^{1,2}. Not surprisingly, as for other brain-behaviour relationships, it has become clear that discrete emotions map onto specific neural circuits in the brain, and this view is now supported by papers^{3,4} on pages 467 and 470 of this issue, along with two reports^{5,6} in *Neuron*. These show that the amygdala, a complex structure within the temporal lobes of the brain, is involved in one form of emotional memory.

Survival depends on the ability of an organism to respond to threat or reward, and to predict the circumstances under which they are likely to occur. Thus, emotional circuits must appraise a situation and initiate adaptive physiological and behavioural responses. To link new predictive stimuli to these adaptive repertoires, emotional circuits must be intimately involved with the formation of memory.

The ability to activate fear circuits in the brain with a neutral stimulus such as a tone (see box, overleaf) has allowed the relevant pathways to be analysed using rats with carefully placed lesions in their brains. Such studies have shown that the amygdala is critical to fear responses and emotional memory^{1,2}. In humans, examination of patients with lesions in the amygdala suggested that the amygdala is involved in fear. Moreover, studies using functional neuroimaging have shown⁷ that the human amygdala is activated in response to visual processing of fearful faces, even when a subject does not consciously realize that he has seen a face. This can be done by presenting a brief fearful stimulus, immediately followed by a longer neutral masking stimulus⁸.

Morris *et al.*³, LaBar *et al.*⁵ and Buchel *et al.*⁶ and have now used functional neuroimaging to investigate human emotional memory resulting from fear conditioning. They show that the amygdala is indeed involved in this form of emotional memory. Moreover, Adolphs *et al.*⁴ report that three patients with complete bilateral damage to the amygdala cannot form accurate social judgements based on facial expression —

Conditions for conditioning

The classical conditioning described by Pavlov over 70 years ago¹¹ is a simple form of associative learning. A neutral stimulus is temporally paired with a stimulus that has a high intrinsic behavioural significance, and is known as the unconditioned stimulus (US). As a result of this pairing, a link is formed between the neutral stimulus and the US. So,

what was previously the neutral stimulus is transformed into a so-called conditioned stimulus (CS). The CS shows a behavioural response that is appropriate to the US.

In fear conditioning, a neutral stimulus (a tone, for example) is given to rats before a naturally aversive stimulus such as a foot shock. After even a single trial, the rat behaves as if the

tone (CS) predicts the time of arrival and magnitude of the shock. Measurable responses include freezing on the spot (a behavioural response) and activation of the sympathetic nervous system (a physiological response). If a CS is presented repeatedly without a reinforcing US, the association is eventually weakened – a process called extinction. **S.E.H.**

a process that presumably depends on the ability to access memories based on previous emotional experience (although, as noted by the authors, this analysis cannot rule out innate bias against certain facial features). These subjects consistently rated faces as more approachable and trustworthy than did subjects with unilateral lesions in the amygdala, or normal controls. The greatest deviation was seen for those faces that the control subjects rated as being the most negative. Thus, the amygdala seems to be required to gain access to socially and emotionally relevant information about danger elicited by human faces.

New functional magnetic resonance imaging (fMRI) techniques, such as those used by LaBar *et al.* and Buchel *et al.*, are well suited to investigating conditioning. They can detect a signal from the brain after a discrete stimulus and, unlike positron emission tomography, without needing successive repetitions of a task. This is important because such repetition might lead to sensitization or habituation to a stimulus⁹, especially in the amygdala, which rapidly desensitizes to presentation of fearful faces⁸.

Buchel *et al.* paired one set of neutral faces with an aversive auditory stimulus (CS+). Compared with unconditioned neutral faces (CS–), the authors found that CS+ induced differential activation of the amygdala and anterior cingulate gyri (areas of cerebral cortex known to be involved in emotion), and other cortical regions. Activation seemed to reflect the learned association between the previously neutral stimulus (a face) and the aversive tone. But whereas the amygdala was quickly desensitized to the CS+, the activated cortical regions were not. LaBar *et al.* studied both acquisition and extinction of conditioned fear. They found that the amygdala and periamygdaloid cortex were activated early during both learning processes, and then became desensitized, although this became statistically significant only during extinction.

Morris *et al.*³ investigated not the acquisition, but the expression of fear conditioning. Using positron emission tomography they monitored neural activity in people who were presented with two angry faces. One of the faces (CS+), through prior conditioning, was associated with a burst of aversive noise; the other (CS–) was not. In half the trials, awareness of the angry face was prevented by backward masking with a neutral face. The authors observed a significant neural response to masked presentations of the conditioned angry face in the right, but not the left, amygdala. Unmasked presentations of the same face produced increased activity in the left – but not the right – amygdala.

Interestingly, using fMRI, LaBar *et al.* found a *right* bias during unmasked acquisition of fear conditioning. But, comparing individual scans, they found marked differences between people in the laterality of amygdala activation during both acquisition and extinction. Although it is clearly premature to speculate about whether one hemisphere or the other may be specialized for

some aspects of fear conditioning, we can conclude that the amygdala is critical in human emotional memory.

Activation of the amygdala seems to be greatest during the early stages of learning^{8,10}. However, the time that it takes to acquire brain images (a matter of seconds to minutes) is a veritable aeon in the nervous system, making interpretation of the current data difficult. Better technologies such as high-field-strength fMRI (which will improve spatial resolution), as well as the combination of fMRI with techniques that provide better temporal resolution, should allow us to correlate investigations in humans with physiological and anatomical work in rodents. The new papers also remind us how subtle the nervous system is – small differences in setting up superficially similar experimental tasks can lead to markedly different neuroimaging results, as seen by the spatial differences in activation of the ‘amygdala’ across the papers. Overall, however, we can look forward to rapid and substantial progress in understanding how emotion is processed in the brain, to develop better approaches for treating mood and anxiety disorders. □

Steven E. Hyman is at the National Institute of Mental Health, National Institutes of Health, 31 Center Drive, Bethesda, Maryland 20892-2475, USA.

e-mail: shyman@helix.nih.gov

1. Rogan, M. T. & LeDoux, J. E. *Cell* **85**, 469–475 (1996).
2. Davis, M. J. *Neuropsychiatr. Clin. Neurosci.* **9**, 382–402 (1997).
3. Morris, J. S., Öhman, A. & Dolan, R. J. *Nature* **393**, 467–470 (1998).
4. Adolphs, R., Tranel, D. & Damasio, A. R. *Nature* **393**, 470–474 (1998).
5. LaBar, K. S., Gatenby, J. C., Gore, J. C., LeDoux, J. E. & Phelps, E. A. *Neuron* **20**, 937–945 (1998).
6. Buchel, C., Morris, J., Dolan, R. J. & Friston, K. J. *Neuron* **20**, 947–957 (1998).
7. Breiter, H. C. *et al.* *Neuron* **17**, 875–887 (1996).
8. Whalen, P. J. *et al.* *J. Neurosci.* **18**, 411–418 (1998).
9. Buckner, R. L. *et al.* *Proc. Natl Acad. Sci. USA* **93**, 14873–14878 (1996).
10. Quirk, G. J., Armony, J. L. & LeDoux, J. E. *Neuron* **19**, 613–624 (1997).
11. Pavlov, I. P. *Conditioned Reflexes* (Dover, New York, 1927).

Planetary science

Making and braking asteroids

Alan W. Harris

There is an old joke about the limitations of physics and its need for mathematical abstraction, the punchline of which runs, “We assume a spherical chicken...”. In trying to model the collisional evolution of asteroids, however, theorists have been unable to avoid assuming a spherical chicken. To create the present asteroid belt, asteroids must have been undergoing collisions throughout the age of the Solar System, but to make tractable analytical calculations of what happens when one asteroid hits another, one is forced into the unrealistic assumption that the bodies are spherical and rigid.

Now at last we can do something better: as reported on page 437 of this issue¹, Asphaug and his co-authors (along with several other research groups) are progressing beyond analytical calculations to detailed numerical models of what happens when irregular, inhomogeneous asteroids suffer high-velocity impacts. Of particular interest are the near-Earth asteroids – not just ‘making’ them, but the possibilities for ‘braking’ one that might chance to be on a collision course with the Earth (the subject of a couple of Hollywood productions, *Deep Impact* and *Armageddon*, which either have been or will shortly be released).

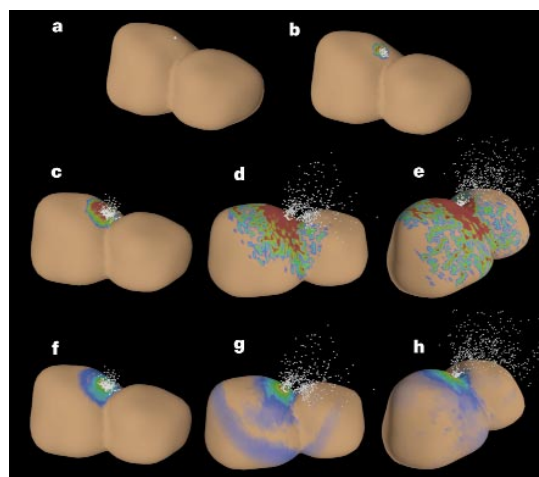


Figure 1 Simulation of an 8-m-radius meteoroid striking the near-Earth-asteroid 4769 Castalia at 5 km s^{-1} . a, The projectile is about to strike the asteroid. b–e, The initially solid asteroid becomes fractured (0.03 s, 0.05 s, 0.14 s and 0.28 s after impact): fully fragmented material — rubble — is red; lesser levels of damage are blue, green and yellow. f–h, Particle velocity at the surface of the asteroid: blue is 10 cm s^{-1} ; green, 100 cm s^{-1} ; red, 1 m s^{-1} ; white is projectile material; escape speed averages about 40 cm s^{-1} . (Further details are available from ref. 9.)

Such detailed modelling has now become possible, partly because we now know the shapes of a few asteroids from radar mapping and spacecraft imaging, but mostly because high-speed supercomputers and advanced hydrodynamic codes are becoming available for non-military applications. Although the 'peace dividend' has yet to pay any obvious cash rebates, it appears to be yielding some research benefits. Over the past few years, similar computations have yielded realistic models of the tidal breakup² of the progenitor of comet Shoemaker–Levy 9 by Jupiter in 1992; models of the impacts and plume formations³ of those fragments as they plunged into the atmosphere of Jupiter in 1994; similar models of the entry into the Earth's atmosphere of giant meteoroids such as the Tunguska bolide⁴ of 1908; and simulations of the formation of the Moon by an enormous glancing collision between some large body and the Earth, early in the Solar System⁵. A rich suite of problems has become ripe for solution with the advent of these powerful computing tools.

An important question about the structure and evolution of asteroids is whether they are monolithic bodies or 'rubble piles' — that is, loose agglomerations of debris fractured by past collisions but not quite dispersed into separate bodies. Asphaug *et al.* conclude that this is likely for many asteroids, but that it may depend critically on the first impact suffered by an initially monolithic body (Fig. 1). Once fracture zones are created, they tend to protect the material on one side from damage due to impacts on the other side, as the shock wave is absorbed, reflected or attenuated by the smaller rocks in the zone. Generally speaking, these effects tend to make the asteroid more resistant to global disruption, and hinder the rate at which further fracturing occurs to convert the entire mass into a 'rubble pile'. A few large pieces can survive.

There is observational evidence that even very small asteroids are rubble piles, from the statistics of spin rates⁶. It seems that no asteroids spin so fast that they would be in a state of tension. Among very small asteroids, the

average rate of spin is fast enough that there appears to be a barrier in the distribution of spins at the rate corresponding to this limit, suggesting that most asteroids indeed have no tensile strength, as would be the case for heavily fractured bodies. On the other hand, rubble pile structure does not help to explain the few very slowly spinning asteroids. 253 Mathilde, for example, imaged in detail by the NEAR spacecraft last year, is one such very slowly spinning body, with a rotation period of around 17 days. The images returned were astonishing for the number and large size of craters revealed, so it is not plausible that Mathilde somehow avoided spin-inducing impacts. However, the very large craters found on Mathilde do indicate a rubble pile structure, because according to the results of Asphaug *et al.* a monolithic body struck hard enough to produce such large craters would probably be blown apart, whereas a 'sandbag' might avoid complete disruption. The low bulk density of Mathilde found by its gravitational perturbations on the NEAR spacecraft also supports the idea that this is a structure with a lot of vacuum in among the rock⁷.

Perhaps the most interesting aspect of this work for the general public is the implication

for defence against asteroid impacts on the Earth. If an asteroid were found to be on a collision course with the Earth, could we avoid it? The front-running technique is to explode a nuclear bomb some distance from the asteroid, vaporizing a thin layer of its surface on one side, and thus giving it a nudge. But most studies⁸ of this process have suffered from the spherical chicken problem, modelling the offending asteroid as a coherent solid body rather than a loose collection of debris. The new work may mean that deflecting an asteroid from a collision course would be more like clearing a landslide off the road than pushing a boulder aside. If the only thing holding the body together is gravity, then one cannot apply an impulsive change in its motion greater than the escape velocity from the surface without disrupting the body into many pieces. This means a kilometre-sized body can be given a change of course of only about a metre per second. Such a small impulse would have to be applied a fair fraction of a year before the projected time of collision in order to accumulate a change of path of a couple of Earth radii. The smaller the object, the smaller the impulse allowed, so the concept of a 'Star Wars' type shield protecting the Earth from imminent impacts is seriously flawed; better to discover asteroids far in advance in an orderly survey, allowing plenty of time to respond. □

Alan W. Harris is in the Earth and Space Sciences Division of the Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California 91109, USA.

e-mail: awharris@lithos.jpl.nasa.gov

1. Asphaug, E., Ostro, S. J., Hudson, R. S., Scheeres, D. J. & Benz, W. *Nature* **393**, 437–440 (1998).
2. Asphaug, E. & Benz, W. *Icarus* **121**, 225–248 (1996).
3. Zahnle, K. & Mac Low, M. *Icarus* **108**, 1–17 (1994).
4. Chyba, C. F., Thomas, P. J. & Zahnle, K. J. *Nature* **361**, 40–44 (1993).
5. Cameron, A. G. W. *Icarus* **126**, 126–137 (1997).
6. Harris, A. W. *Lunar Planet. Sci.* **27**, 493–494 (1996).
7. Yeomans, D. K. *et al.* *Science* **278**, 2106–2109 (1997).
8. Ahrens, T. J. & Harris, A. W. *Nature* **360**, 429–433 (1992).
9. <http://www.es.ucsc.edu/~asphaug>

Desert ecology

Life in the upper crust

Peter D. Moore

The surfaces of deserts are not as bare as they may appear — they get their crusty nature from the dense growth of many organisms in the upper millimetre of dry soil. These organisms not only influence the behaviour of water in the soil, but they also contribute substantially to nitrogen and carbon fixation. Research by O. L. Lange and his colleagues over the past few years, including their latest report in *Functional Ecology*¹, has shown that deserts differ in their crustal communities. And we now know that these highly active layers of desert soils contain productive species that are

finely tuned in their physiology to the nature and the timing of their water supply, whether in the form of dew or rain showers.

Desert crusts consist of a tangled mesh of various types of organism that bind, in a mucilaginous, fragile concretion, the fine-textured inorganic particles that form their matrix. Their living components include algae, mosses (including protonemal forms), lichens, fungi, bacteria and cyanobacteria. These organisms occupy just the upper crust, because many of them need light for their photosynthetic activities and also because water is briefly available in these



Figure 1 Early morning in the Namib — the optimal time for crustal photosynthesis.

superficial layers, derived from dew, fog and showers. The ecological value of the layer in protecting the soils from wind erosion and as an absorptive organ for water is generally appreciated, as is its role in providing germination sites for the seeds of flowering plants².

The presence of free-living, nitrogen-fixing cyanobacteria in the upper layer, together with lichens that contain cyanobacterial symbionts, suggests a pathway by which atmospheric nitrogen enters the desert ecosystem. Moreover, the extensive cover of even a thin layer of photosynthetic organisms is an important source of carbon fixation in an environment where primary production is low. In their studies of the Namib Desert of southern Africa (Fig. 1), Lange and colleagues have shown³ that the chlorophyll density of the microbial mat is around 200–500 mg m⁻² — of the same order as in the leaf of a higher plant. So the crust has considerable photosynthetic potential, although this is limited by the need for hydration before it can be exploited.

In the Negev Desert of Israel, Lange and his team showed⁴ that they could induce maximal photosynthesis from the crust by moistening it with the equivalent of just 0.2–0.3 mm of precipitation. These photosynthetic rates were around 20% of those achieved by local desert shrubs, such as *Zygophyllum dumosum*, illustrating the ability of the crust to act as a carbon sink. In the Negev, as in the Namib, moisture mainly comes in the form of condensation — as dew and fog during the cold night. So, although these moisture levels are often achieved, the rising sun soon desiccates the productive crust into a dormant state, and crustal photosynthesis is limited to an early-morning window when moisture and light briefly coincide.

Now that general taxonomic surveys of crust communities and their physiological properties have been carried out, questions of geographical and ecological variations are being tackled. For example, not all deserts depend on condensation. The intermontane deserts of the western United States

are watered by occasional showers or storms rather than by fog and dew — do their crusts have a different species composition and physiology? To answer these questions, Lange *et al.* have done their latest work¹ in southern Utah.

One evident taxonomic feature of the Utah desert is that, although there are many lichens with green algae as their symbionts (as in the Negev and Namib deserts), the most abundant crustal constituent is a gelatinous lichen, *Collema tenax*, with a cyanobacterial symbiont. This lichen requires a greater water supply (equivalent to over 1 mm of precipitation) than do the green-algal lichens to attain its maximum photosynthetic rate. But, it then shows the highest photosynthetic capacity of any arid-zone soil lichen yet studied. Moreover, *C. tenax* can maintain photosynthetic activity over a wide range of temperatures, from near freezing to over 40 °C, with an optimum around 30 °C. It also requires a high light intensity for maximal photosynthesis, and it has a high light-compensation point (the minimum light intensity at which it can break even in terms of carbon balance). All of these are features that plant physiologists would associate with a 'sun plant'.

Collema tenax has a distinctive physiology that equips it well for the water regime in Utah, where erratic showers bring larger quantities of rain at any given time than are experienced in the 'fog' deserts. The gelatinous nature of its thallus means that *C. tenax* serves as a reservoir, maintaining its hydrated state long enough to provide for extended photosynthesis even in the hot conditions, when the light is high, in the middle of the desert day. Lichens with cyanobacterial symbionts are already known to need more water to set photosynthesis in motion than lichens with algal symbionts. So, perhaps the features described for *C. tenax* will be of wide taxonomic, physiological and biogeographical significance.

Recognizing that such niche differentiation exists among the primary producers in desert crusts is an important step towards appreciating the complexity of crustal micro-ecology. Because these soil features are so important in the water relations and nutrient cycling of desert soils, and because we know that they are sensitive to disturbance by the trampling pressures of humans, domestic animals and motor vehicles, further understanding of crustal ecology should soon provide a basis for conservation of desert crusts. □

Peter D. Moore is in the Division of Life Sciences, King's College, Campden Hill Road, London W8 7AH, UK.

1. Lange, O. L., Belnap, J. & Reichenberger, H. *Funct. Ecol.* **12**, 195–202 (1998).
2. West, N. E. *Adv. Ecol. Res.* **20**, 179–223 (1990).
3. Lange, O. L., Meyer, A., Zellner, H. & Heber, U. *Funct. Ecol.* **8**, 253–264 (1994).
4. Lange, O. L. *et al. Funct. Ecol.* **6**, 519–527 (1992).

Daedalus

Plane and fancy

The synthesis of nylon is a well-known demonstration. One component is dissolved in water, and a second one in an immiscible solvent. On mixing the two, a nylon film forms at the liquid interface, renewed as fast as it is pulled from the mixture.

In principle, the interfacial film should be a perfect monolayer separating the two reagents. Sadly, nylon with its chain-molecules is very permeable. The reagents can still get at each other; they react further to thicken the film. A polymer which cross-linked into a true sheet molecule would be a far better barrier. So Daedalus is devising two-solution syntheses of dense, cross-linked, water-repellent polymers with fluorocarbon or organometallic side-chains, excellent barriers to their precursor reagents. At the solution interface a true monolayer film will form, regenerated as fast as it can be hauled out of the bath.

Monolayer film will be quite invisible. Furthermore, being essentially all surface, its two faces will absorb almost its own weight of air. Unless it is chemically non-inflammable, it will be highly explosive. So it will be quite tricky to handle. On the other hand, it will be immensely strong for its weight, the highest of high-tensile films. It will also be the ultimate electrical capacitor, holding a vast charge-density per unit weight. Monolayer film will make amazing new electrometers, electrostatic speakers, balloon fabrics, and so on.

A wider range of products could be made by bringing many monolayers together into a loosely bonded multilayer. The resulting macroscopic film would be immensely tough. A crack could not propagate easily from layer to layer, so the film should sum the tenacity of its individual monolayers. Even better, Daedalus plans to make multilayers of two or more types of film, interleaved into a multiple laminate structure. If one of the layers were organometallic, the resulting bulk 'superlattice' structure might give a transparent magnet, a high-temperature superconductor, an interference filter, a specular X-ray mirror, a large-scale quantum-tunnelling device or some other supermolecular marvel.

A simpler sandwich structure could have just three films: a central crumpled one bonded to two flat monolayer faces. It would be a wonderful heat insulator. Being transparent, indeed invisible, it could not radiate. Applications to Emperors' new winter clothing should follow.

David Jones

Vesalius's veracity

The sixteenth-century anatomist Vesalius debunked many of the doctrines of the Greek physician Galen. Vesalius chose his illustrations carefully to act as powerful tools in proving the accuracy of his scientific observations.

Martin Kemp

Showing the tools of the trade was a long-standing practice in the illustration of surgical texts, not least in the manuscript tradition that looked towards Islamic sources. With the rise of naturalistic illustration, particularly in the great picture-books of the human and natural sciences in the sixteenth century, the equipment could be displayed with a new conviction. The instructional role of such representations seems obvious. But is their function and meaning as simple as this?

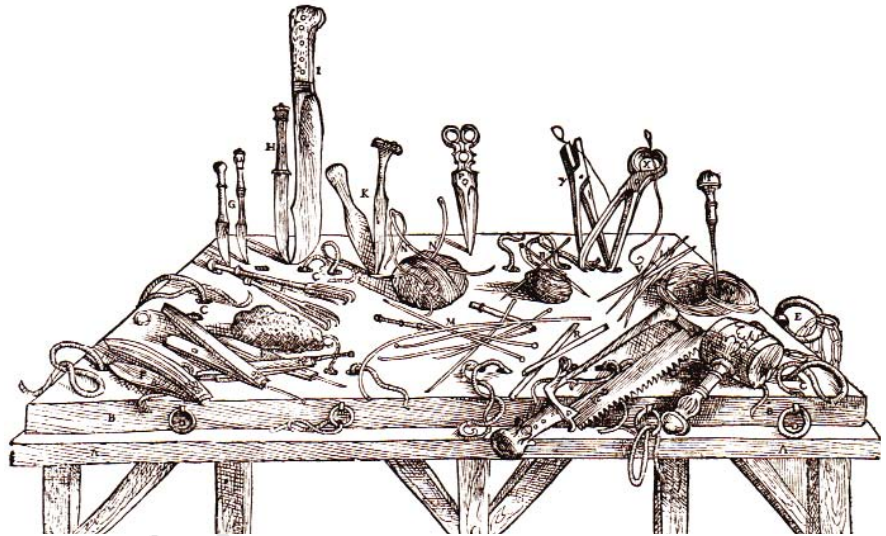
When the anatomist Andreas Vesalius of Brussels displayed in his book an array of dissecting instruments — many of which were common or garden in the literal sense — he was demonstrating to the reader “everything that could be used in the conduct of dissections or an entire anatomy”. The implication is that aspiring dissectors would be fully kitted out if they could lay their hands on everything in the woodcut.

But Vesalius's large and luxurious volume, *De Humani Corporis Fabrica*, dedicated in 1543 to the Holy Roman Emperor Charles V, can hardly be considered as a stock handbook aimed at tirois or even the jobbing surgeon.

The level of knowledge in the *Fabrica* went far beyond the rude empirical procedures needed by a field surgeon or even the kind of court employee Vesalius was aspiring to become. The *Fabrica* was more in the nature of philosophical treatise on the architectural magnificence of the human body, bearing witness to Vesalius's heroic excavation of the inner truths of its fabric.

The illustration of the tools is just one move in a series of visual strategies to underline the veracity of Vesalius's representations. He is at pains to stress that his knowledge is literally first-hand; that is to say he had descended from the aloof heights of the professorial cathedra to undertake the cutting on his own account.

This direct intervention extends beyond anatomies of the dead. He tells us that the board bearing the instruments is “such as we employ in vivisections, resting on a table”. This board, illustrated again on two occasions in the *Fabrica* — once with an unfortunate pig tethered to its rings and apertures, and also within an illuminated letter ‘Q’ with eager *putti* performing the vivisection — makes open reference to the experiments by



Vesalius's “Tools for Dissection” from *De Humani Corporis Fabrica*, Basel, 1543.

Galen, the Alexandrian philosopher and doctor.

Vesalius is at once aspiring to surpass Galen in first-hand knowledge and setting himself up as a second Galen — following his great predecessor's principles of deducing function rigorously from observed form.

In his famous series of musclemen, heroically performing their myological striptease in grand landscapes, Vesalius repeatedly emphasizes the physical reality of his procedures, telling us, for instance, how the figures were suspended by ropes. Less well known is his innovative use of visual proof through mechanical analogies, as in his illustrations of the metal hinge of a window shutter, and his neat demonstration of the restraining role of the transverse ligament in the ankle.

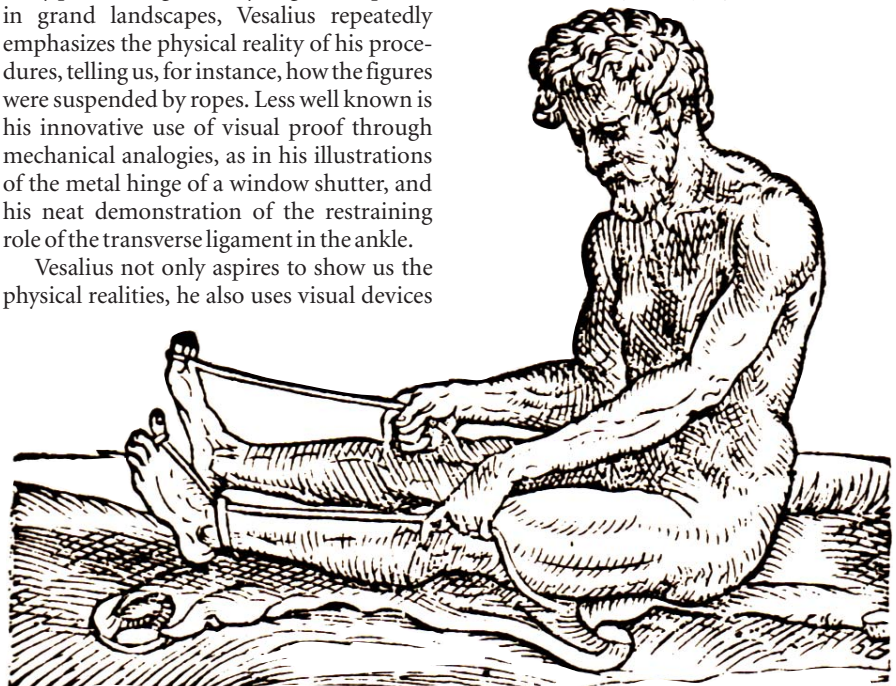
Vesalius not only aspires to show us the physical realities, he also uses visual devices

to convince the reader of the physical truth of his observations.

Among those who emulated such devices were Gottfried Bidloo, who in 1685 showed a fly on a dissection, and William Hunter (1774), who depicted a reflected window on a moist membrane over a fetus.

Martin Kemp is in the Department of the History of Art, University of Oxford, 35 Beaumont Street, Oxford OX1 2PG, UK.

e-mail: martin.kemp@trinity.oxford.ac.uk



Vesalius's “Demonstration of the Need for a Transverse Ligament in the Ankle” from *De Humani Corporis Fabrica*.

Dragon fish see using chlorophyll

Most deep-sea fish have visual pigments that are most sensitive to wavelengths around 460–490 nm, the intensity maxima of both conventional blue bioluminescence and dim residual sunlight¹. The predatory deep-sea dragon fish *Malacosteus niger*, the closely related *Aristostomias* sp. and *Pachystomias microdon* can, in addition to blue bioluminescence, also emit far-red light from suborbital photophores², which is invisible to other deep-sea animals. Whereas *Aristostomias* sp. enhances its long-wavelength sensitivity using visual pigments that are unusually red sensitive³, we now report that *M. niger* attains the same result using a derivative of chlorophyll as a photosensitizer.

Aristostomias tittmanni has at least three visual pigments, based on two opsins, with in its retina (absorbance peaks at about 531, 548 and 588 nm)³. For the dragon fish, *M. niger*, we isolated outer segments of the photoreceptor by suspending them in 20% sucrose solution, ensuring that minimal disruption was caused to the photoreceptor membranes. Only two visual pigments (absorbance peaks at 517 and 542 nm) were identified by partial-bleaching spectrophotometry of both detergent extracts (Fig. 1a,b) and fresh outer-segment suspensions (data not shown).

Spectrophotometry identified these two pigments as a rhodopsin/porphyropsin pair based on a single opsin bound to retinal or 3,4-dehydroretinal. We verified that there was only one opsin in *M. niger* by using degenerate primers, based on the sequence of teleost fish genes encoding rod opsins, to amplify a single opsin gene from *M. niger* genomic DNA (Genbank accession no. AJ224691). That this gene encodes a rod opsin was confirmed by cladistic analysis and by the absence of introns, a feature that seems to be unique to the rod opsin genes in teleosts⁴. A 259-base-pair 5' fragment of this gene was used to probe at low stringency a Southern blot of *M. niger* DNA digested with restriction enzyme. Only a single fragment hybridized from each digest.

M. niger thus lacks the long-wavelength opsin that in *A. tittmanni* allows extreme red sensitivity, but the outer segments of the *M. niger* photoreceptor do contain a photo-stable pigment complex with a long-wavelength absorption peak (666.8 nm) close to the emission maximum of its red bioluminescence (705 nm; Fig. 2b)^{5,6}. This complex could act as a photosensitizer, absorbing longer wavelengths and transferring the energy so derived to visual pigments that are sensitive to shorter wavelengths⁶, as demonstrated by photobleaching of *M. niger* retinal extracts (Fig. 1c) and suspensions of the rod outer segment (data not shown).

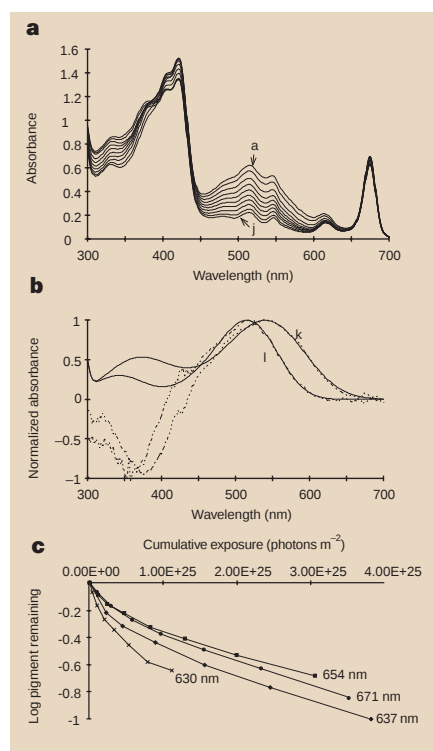


Figure 1 Analysis of *Malacosteus niger* visual pigments. **a**, Absorbance spectra of a retinal extract following exposure to monochromatic light. In order of decreasing absorbance at 500 nm: curve a: initial measurement after the addition of hydroxylamine; b–h: following 5, 15, 30, 50, 80, 120 and 180 min of 671-nm illumination; i, j: following 2 and 12 min of 501-nm illumination. **b**, Normalized difference spectra constructed using the curves in **a**; (k) a, b; (l) i, j. These best fit to templates of a porphyropsin ($\lambda_{\text{max}} = 540$ nm) and a rhodopsin ($\lambda_{\text{max}} = 515$ nm). Data derived from seven retinal extracts show these pigments to have average λ_{max} values of 542 and 517 nm. **c**, Effectiveness of light of different wavelengths at bleaching *M. niger* visual pigment extracts.

Exposure to 671 nm, close to the absorption peak of the presumed photosensitizing pigment complex (Fig. 2), is more effective at bleaching the visual pigments than a 654-nm stimulus. If the stimulating light was bleaching the visual pigments directly, the 654-nm stimulus would be more effective than that at 671 nm.

To identify the putative photosensitizer, we analysed suspensions of retinal cells in 20% sucrose. The fluorescence excitation and emission spectra (Fig. 2a) suggest a magnesium-free chlorophyll derivative such as pheophorbide *a* or pyropheophorbide *a*. When homogenized retinal suspensions were extracted in hexane/methanol, the photosensitizing pigment was entirely in the methanol phase, indicating that it possessed a free propionic-acid residue.

Thin-layer chromatography of the chlo-

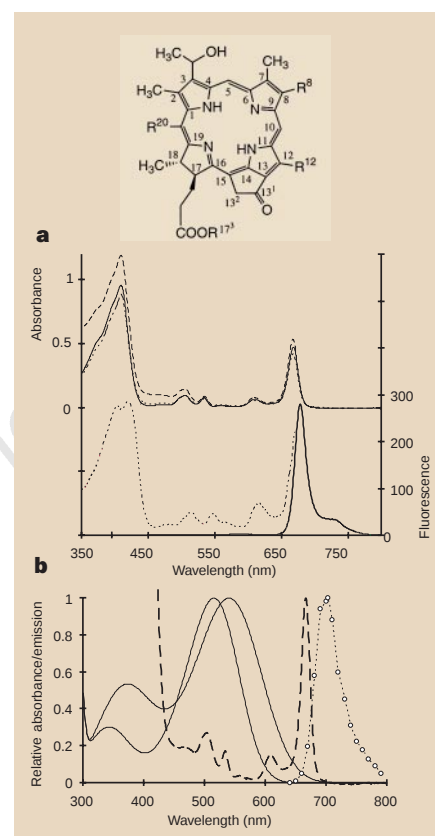


Figure 2 Analysis of *Malacosteus niger* photosensitizing pigment. **a**, Lower spectra: fluorescent excitation (dotted line) and emission (solid line, excitation at 418 nm) spectra of unpurified *M. niger* retinal cell suspension. The excitation spectrum (roughly equivalent to the absorption spectrum of the pigment) was produced by measuring the emission at the peak wavelength of 678 nm. Upper spectra: absorbance spectrum of purified *M. niger* photosensitizing pigment(s) (dashed line); spectra of standard samples of pheophorbide *a* (solid line) and pyropheophorbide *a* (dotted line) prepared from pure chlorophyll *a*. Inset, the structure of known chlorobium pheophorbides: R^8 is Et, Pr^1 , Bu^1 or Pn^{neo} , R^{12} is Me or Et; R^{20} is Me or H; R^{173} is Me. **b**, Absorbance spectra of *M. niger* photoreceptor pigments shown in relation to the emission spectrum of the long-wave suborbital photophore. Solid lines, two visual pigment templates: a rhodopsin with λ_{max} 515 nm and a porphyropsin with λ_{max} 540 nm; dashed line, absorption spectrum of the presumed photosensitizer (a purified diethyl ether extract of a retinal suspension); circles, far-red bioluminescence².

reform extract of the methanol phase gave only one spot for the methyl esters of the photosensitizing pigment(s) that migrated at the same speed as the reference compounds, methyl pheophorbide *a* and pyropheophorbide *a*, prepared from pure chlorophyll *a* (ref. 7).

The ultraviolet-visible absorption spec-

trum of the diethyl ether extract of this unknown compound(s) (Fig. 2a) was similar to that of standard samples of pheophorbide *a* and pyropheophorbide *a*, and virtually identical with the spectrum of the methyl ester of chlorobium (Cbm) pheophorbide 650, fraction 2. The electron-impact mass spectrum for the methyl esters of the unknown compound(s) suggests that they are Cbm pheophorbides (a mixture of defarnesylated and demethylated derivatives of Cbm chlorophylls, series 650 and 660). High-intensity signals corresponding to the expected main fragments from Cbm pheophorbide 650 methyl esters suggest that the Cbm chlorophyll 650 derivatives predominate.

Energy transfer from the photosensitizer to the visual pigment may involve inter-system crossing to generate an excited triplet state (T_1) of the pheophorbide, which could then transfer energy to the ground state of the visual pigment's 11-*cis* chromophore to generate its triplet state. Such a triplet state of the visual pigment chromophore has been proposed as a normal intermediate of photoisomerization⁸. Energy transfer from the triplet state of the pheophorbide *a* derivative ($E_r = 30.7 \text{ kcal mol}^{-1}$)^{9,10} to the ground state of the 11-*cis* chromophore ($E_r < 38 \text{ kcal mol}^{-1}$) might be endothermic, but the long lifetime (1 ms) of the pheophorbide triplet state and the high quantum yield of its formations make the triplet photosensitization mechanism highly probable.

R. H. Douglas

Applied Vision Research Centre, Department of Optometry and Visual Science, City University, 311–321 Goswell Road, London EC1V 7DD, UK
e-mail: r.h.douglas@city.ac.uk

J. C. Partridge

School of Biological Sciences, University of Bristol, Woodland Road, Bristol BS8 1UG, UK

K. Dulai, D. Hunt

Institute of Ophthalmology, University of London, Bath Street, London EC1V 9EL, UK

C. W. Mullineaux

Department of Biology, University College London, Gower Street, London WC1E 6BT, UK

A. Y. Tauber, P. H. Hynninen

Department of Chemistry, University of Helsinki, PO Box 55, 00014 Helsinki, Finland

- Douglas, R. H., Partridge, J. C. & Hope, A. J. *J. Comp. Physiol. A* **177**, 111–122 (1984).
- Widder, E. A., Latz, M. I., Herring, P. J. & Case, J. F. *Science* **225**, 512–514 (1995).
- Partridge, J. C. & Douglas, R. H. *Nature* **375**, 21–22 (1995).
- Fitzgibbon, J. *et al. Gene* **164**, 273–277 (1995).
- Partridge, J. C., Shand, J., Archer, S. N., Lythgoe, J. N. & van Groningen-Luyben, W. A. H. M. *J. Comp. Physiol. A* **164**, 513–529 (1989).
- Bowmaker, J. K., Dartnall, H. J. A. & Herring, P. J. *J. Comp. Physiol. A* **163**, 685–698 (1988).
- Hynninen, P. H. in *Chlorophylls* (ed. Scheer, H.) 145–209 (CRC Press, Boca Raton, 1991).
- Jensen, N.-H., Willbrandt, R. & Bensasson, R. V. *J. Am. Chem. Soc.* **111**, 7877–7888 (1989).
- Seely, G. R. *Photochem. Photobiol.* **27**, 639–654 (1978).
- Krasnovsky, A. A. Jr *et al. Photochem. Photobiol. B: Biol. Sci.* **245**–254 (1990).

Vision and attention: the role of training

What happens to visual experience in the absence of visual attention? Does lack of attention render us effectively blind¹, or is there a significant residual experience^{2–4}? Here I show that the surprising results of a recent study¹ were due not to the novel way in which attention was controlled, but simply to the use of novice rather than expert observers. So the evidence remains strong that some aspects of visual experience are essentially independent of attention.

Joseph *et al.*¹ reported that observers are unable to detect a simple feature difference ('popout') when attention is taken up elsewhere in the display. This is at odds with previous reports^{2–4} and a large body of work on preattentive, parallel processes in vision^{5–7}.

Methodologically, the new study departs from earlier efforts by controlling attention with the help of an 'attentional blink'. This is the lapse in perception that occurs, for example, when a predesignated target letter is recognized in a stream of letters appearing rapidly one after the other. Joseph *et al.* argue that attentional blink removes attention more effectively than the concurrent-task method used in previous studies^{2–4}.

However, a second methodological difference that they do not emphasize is the use of novice (those who have performed only 480 trials) rather than expert (many thousands of trials) observers.

To ascertain which of these factors was responsible for the outcome, I investigated three observer populations with the attentional blink method of Joseph *et al.* (Fig. 1a): novice observers who had no prior experience with tachistoscopic displays; trained observers who practised the experiment for thousands of trials before data collection; and expert observers with extensive prior experience with tachistoscopic displays but no practice with the present experiment (or other concurrent task experiments).

The results are shown in Fig. 1c, e, g. 'Popout' detection during the attentional blink was at or near chance in novices (Fig. 1c) but close to normal performance levels in trained and expert observers (Fig. 1e, g). As well as popout detection, I investigated the attentionally more demanding 'T/L' discrimination. For this task, the attentional blink severely impaired performance in all three observer populations (Fig. 1b, d, f, h).

It seems, therefore, that the attentional blink method produces the same pattern of results as the concurrent-task method of earlier studies^{2–4}: in trained or expert observers, popout detection is largely unaffected but T/L discrimination is severely

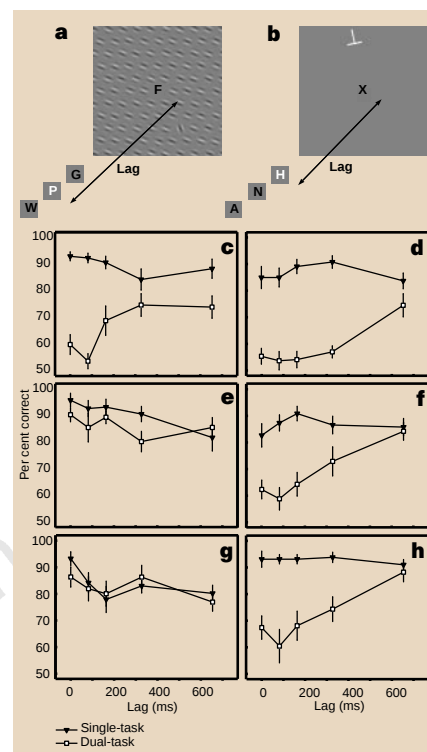


Figure 1 'Attentional blink' experiments¹. Subjects fixate on a rapidly presented stream of black letters (12 per s) and attempt to identify the one white letter in the stream ('RSVP task', chance performance 3.8% correct). Some time after the white letter (lag time 0, 82, 164, 327 or 655 ms), a second target appears at a more peripheral location (3–4 deg eccentricity). **a**, Second target is a uniquely oriented Gabor element and observers report 'present/absent' ('popout detection'). **b**, Second target is a single white letter (rotated T or L) and observers report 'T/L' ('T/L discrimination'). **c–h**, Peripheral task performance as a function of lag time when the peripheral task is carried out alone (single-task, all trials) and when it is carried out concurrently with the RSVP task (dual-task, trials in which RSVP response was correct, 75–81% of all trials). Average performance $\pm$ s.e. shown for all observers. **c, e, g**, Popout detection as the peripheral task. **d, f, h**, T/L discrimination as the peripheral task. **c, d**, Novice observers: six observers in first encounter with tachistoscopic displays. **e, f**, Trained observers: two novice observers after extensive training (4 days/2,560–2,880 trials on **a** and, subsequently, 5–6 days/4,080–4,560 trials on **b**). **g, h**, Expert observers: three observers experienced with tachistoscopic displays but without training on this, or any other dual-task experiment.

impaired (Fig. 1e, f, g, h and ref. 4), whereas in novice observers, both tasks are disrupted to the point of being performed at or near chance (Fig. 1c, d; ref. 1, and J.B., unpublished observations).

So trained or expert observers consistently show large differences in the degree to which various tasks depend on attention, whereas in novice observers the performance of all secondary tasks seems to be equally poor. Clearly, Joseph and colleagues

obtained their result because they relied on novice observers, not because attentional blink removes attention any more or less effectively than a concurrent task.

Why are novice observers different? It is often thought that training reduces the attentional demands of visual tasks. However, this would not explain why some tasks (for example, popout detection; Fig. 1c, e) lose so much and others (such as T/L discrimination; Fig. 1d, f) lose so little of their attentional demand. It also fails to explain why prior experience with tachistoscopic displays reduces attentional demands at least as effectively as specific training of a particular task (Fig. 1e, g).

Perhaps a more likely reason is that observers generally fare poorly with unexpected or unfamiliar stimuli^{8,9}. Stimuli that are tachistoscopically presented and less than fully attended to do not occur in everyday vision, and to novice observers the subjective appearance of such stimuli may be unfamiliar. Either training or experience may enhance awareness of such stimuli, without any real change in the attentional demands of associated visual tasks.

The main point, however, is that no attentional load seems high enough to prevent performance of certain basic visual tasks by trained or expert observers. In such observers at least, there does seem to be a 'direct route from preattentive processing to perceptual report'¹. Furthermore, the disparate attentional requirements of various visual tasks are likely to reflect differences in the underlying neural substrate. Thus, experiments with trained or expert observers remain an excellent tool for linking perception and neural substrate.

Jochen Braun

Biology 139-74, Caltech, Pasadena,
California 91125, USA
e-mail: achim@klab.caltech.edu

1. Joseph, J. S., Chun, M. M. & Nakayama, K. *Nature* **387**, 805–807 (1997).
2. Braun, J. & Sagi, D. *Percept. Psychophys.* **48**, 45–48 (1990).
3. Braun, J. J. *Neuroscience* **14**, 554–567 (1994).
4. Braun, J. & Julesz, B. *Percept. Psychophys.* **60**, 1–23 (1998).
5. Julesz, B. *Rev. Mod. Phys.* **63**, 735–772 (1991).
6. Treisman, A. *Attention: Selection, Awareness, Control* (eds Baddeley, A. & Weiskrantz, L.) 1–35 (Clarendon, Oxford, 1993).
7. Watt, R. *Cogn. Neurosci.* **1**, 137–139 (1991).
8. Mack, A., Tang, B., Tuma, R., Kahn, S. & Rock, I. *Cogn. Psychol.* **24**, 475–501 (1992).
9. Rock, I., Linnett, C. M., Grant, P. & Mack, A. *Cogn. Psychol.* **24**, 502–534 (1992).

Joseph et al. reply — We are grateful that Braun has replicated our findings, showing that orientation oddball detection in visual search is severely compromised when paired with an attentional blink task¹. This confirms our conclusion that tasks that have generally been considered to be 'pre-attentive', exhibiting parallel search and rapid discrimination performance, actually do require attention. These data favour a unitary large-scale architecture for the visual

system, in which all the output of an unlimited-capacity stage (preattentive processing) must pass through a limited-capacity selection stage (attentional selection) before explicit detection. Braun then presents two other experiments which show that such a decrement in performance is reduced with increased training and concludes: "...In such observers at least, there does seem to be a 'direct route from preattentive processing to perceptual report'".

We are surprised by such a statement because it implies a fundamental change in the global visuo-cognitive architecture with practice. Once the novice subject becomes practised, we are meant to suppose, there is then a "direct route from preattentive processing to perceptual report". In fact, there are a large number of degrees of freedom in the local circuitry of preattentive processing that can lead to increased stimulus encoding efficiency, resulting in a reduced attentional load and eventual elimination of task interference with practice. This all occurs within a unitary architecture, without resorting to Braun's scheme for a global architectural change.

To illustrate how a simpler unitary architecture can account for the practice effects observed by Braun and ourselves (page 806 of ref. 1), we refer to Norman and Bobrow's² treatment of dual-task studies. Performance in any visual task requires a certain amount of resources (attention), as well as a sufficient stimulus strength to compete with noise in the visual system. Task performance can thus be limited by either of these factors.

Hence, task performance improves with increased resource allocation to the task (Fig. 1a), until resource allocation is no longer limiting; the 'single-task' performance level is then obtained. This is shown in Fig. 1a for a variety of visual tasks of varying 'difficulty'. A relatively 'easy' task has a curve that rises sharply and quickly reaches its maximum performance, while a 'hard' task rises slowly and requires nearly all the resources to achieve its maximum. Practice increases the efficiency with which the stimulus is encoded, reducing its 'attentional load' (the amount of attention needed to reach maximum performance). This is illustrated in Fig. 1b and c as a contraction of each curve to the left after extensive practice.

To demonstrate how these ideas could account for practice effects within a unitary architecture, assume that our RSVP task withdraws a significant amount of attention, lowering the amount available from the full capacity, F , to a much smaller remaining capacity, R . This is illustrated together with the curves for both novice and highly trained observers engaged in a popout or a peripheral L/T discrimination (Fig. 1b, c).

Because the popout's curve is relatively steep in its resource-limited range, practice

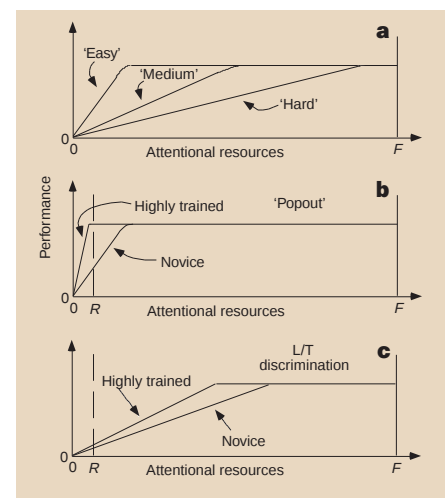


Figure 1 a, Hypothetical task performance as a function of allocated attentional resources for three levels of task 'difficulty'. Maximum performance in all tasks is achieved when the full attentional capacity F is allocated. **b**, Additionally imposing an RSVP task reduces the attentional resources available for a 'popout' task, leaving a small remaining capacity, R , thus severely limiting performance for the novice but not for highly trained observers. **c**, Imposing an RSVP task in addition to an L/T letter discrimination reduces available resources similarly but does not lead to significantly altered performance when comparing novice and highly trained observers.

has a dramatic effect, essentially eliminating the task interference (Fig. 1b). For the more demanding L/T discrimination task, however, the effect of practice is hardly noticeable (Fig. 1c; observe the differences between the two curves for each task). Regarding Braun's 'expert' observers, extensive practice in tasks with briefly presented stimuli is likely to increase coding efficiency for rapid stimuli in general, again resulting in a reduced attentional load and the observed pattern of results. If we accept this explanatory framework, Braun's changeover from a unitary to a dichotomous architecture with practice becomes unnecessary.

As such, the pattern of practice effects obtained by Braun and ourselves is fully expected within a unitary architecture for the visual system, in which all visual information is required to pass through a limited-capacity stage before it can be explicitly detected.

Julian S. Joseph

Department of Psychology, University of Nevada,
Reno, Nevada 89557, USA

Marvin M. Chun

Department of Psychology, Yale University,
New Haven, Connecticut 06520, USA

Ken Nakayama

Department of Psychology, Harvard University,
Cambridge, Massachusetts 02138, USA

1. Joseph, J. S., Chun, M. M. & Nakayama, K. *Nature* **387**, 805–807 (1997).
2. Norman, D. A. & Bobrow, D. G. *Cogn. Psychol.* **7**, 44–64 (1975).

Marine phosphorus is selectively remineralized

Phosphorus is a vital nutrient of the world's oceans^{1,2}, where in vast regions it is associated with dissolved organic matter (DOM) in surface waters^{3,4}. We have characterized the major compound classes of high-molecular-weight marine dissolved organic phosphorus, phosphorus esters and phosphonates, by using tangential-flow ultrafiltration and phosphorus-31 nuclear magnetic resonance (³¹P NMR). We find that the composition and abundance of organic phosphorus in DOM differ significantly from the values in fresh organic matter, indicating that dissolved organic phosphorus (DOP) is preferentially remineralized from DOM.

It has become feasible, with the advent of tangential-flow ultrafiltration, to concentrate and isolate sufficient dissolved and particulate organic matter from natural waters for NMR spectroscopic analysis^{5,6}. Tangential-flow ultrafiltration of sea water reproducibly isolates a greater fraction of DOM than other methods, so ultrafiltered isolates represent the largest fractions of marine DOM ever characterized⁵⁻⁷.

In this study, 19–38% of total DOM was

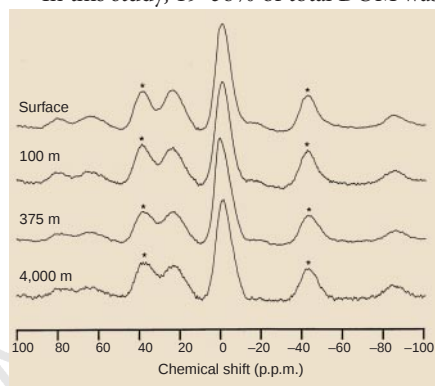


Figure 1 Phosphorus-31 cross-polarized magic-angle spinning NMR spectra of organic matter in sea water. Dried high-molecular-weight DOM isolated from ~1,000 litres of sea water from the surface, 100 m (chlorophyll maximum), 375 m (oxygen minimum) and 4,000 m in the Pacific Ocean (12°S, 134°W)⁶. Spectra were collected on a Chemagnetics CMX300 spectrometer at a frequency of 121.3 MHz and a spin rate of 5 kHz. Chemical shifts are referenced to phosphoric acid. Spectra were acquired using a pulse width of 4.4 μs and a pulse delay of 500 ms. Contact time was 1 ms. Data for surface and 100-m samples were collected over 20,000 transients, and data for 375- and 4,000-m samples were collected over 100,000 transients. Line-broadening of 50 Hz was applied to all samples. Asterisks denote spinning sidebands. NMR experiments with organic P standard compounds were performed to ensure that changes in peak area are linearly related to changes in abundance of the P compounds.

Table 1 Sample depth and concentration of DOC, DON and DOP, and C/N/P ratios

Sample	DOC	DON	DOP	C/N/P
Surface	22.2	1.33	0.090	247/15/1
100 m	22.2	1.42	0.083	267/17/1
375 m	10.6	0.63	0.033	321/19/1
4,000 m	8.08	0.45	0.015	539/30/1

Concentrations of ultrafiltered dissolved organic carbon (DOC), nitrogen (DON) and phosphorus (DOP) are expressed as micromolar; C/N/P values are the calculated atom ratios.

recovered by ultrafiltration⁶. All the ultrafiltered DOM isolates characterized here are in the high-molecular-weight or colloidal fraction (1–100-nm size range)⁶, which is a significant source of carbon and energy for marine bacteria⁸.

The ³¹P NMR spectrum of surface-water DOM indicates that DOP is dominated by two major classes of compound, phosphorus esters and phosphonates (Fig. 1). The most prominent feature of all spectra is the large peak centred at a chemical shift of 0 p.p.m., attributed to phosphorus esters. In marine plankton, phosphorus esters are present as monoester and diester phosphates in macromolecules such as nucleic acids and membrane phospholipids⁹. The smaller peak at 25 p.p.m. indicates the presence of phosphonates, a group of compounds containing a C–P bond. Phosphonates are often associated with phosphoproteins¹⁰ and membrane phosphonolipids⁹. Phosphonolipids in outer-membrane structures may protect cells from enzymatic degradation or provide extra rigidity^{11,12}.

About 3% of total phosphorus in a natural assemblage of plankton was attributed to phosphonates¹³. Although phosphonates comprise only a few per cent of phosphorus in plankton, integration of NMR peak areas reveals a surprisingly high abundance of phosphonates in high-molecular-weight DOP (about 25%). The relatively high proportion of phosphonates in surface DOP indicates preferential use of phosphorus esters relative to phosphonates during organic matter decomposition.

In sea water, some dissolved phosphorus esters are a rapidly used source of phosphorus whereas phosphonates are considered to be chemically stable compounds, more resistant to hydrolysis and bacterial degradation⁹. Given the higher proportion of phosphonates in surface DOP relative to marine plankton, regeneration must initially be a selective process, with reactive phosphorus esters as a preferred substrate for microorganisms.

Although the absolute concentration of high-molecular-weight DOP decreases from 90 nM in surface waters to 15 nM in deep waters (Table 1), ³¹P NMR spectra of DOP show phosphorus esters and phosphonates in unchanging proportions throughout the water column (Fig. 1), indicating that phosphorus esters and phosphonates are used at equivalent rates in deeper waters. Apparent differences in the reactivity of phosphorus esters in surface compared with deeper

waters may result from phosphorus esters occurring in a variety of biochemical classes of differing reactivities. Reactive phosphorus esters are used rapidly in surface waters, resulting in enrichment of deep waters in less reactive DOP compounds. The persistence of phosphorus esters and phosphonates from surface to deep waters suggests that less reactive fractions of DOP consist of common biochemical structures that are probably produced in surface waters.

The regeneration of nutrients in sea water is often compared to the Redfield ratio¹, which approximates the average composition of marine planktonic organisms (ratio of carbon/nitrogen/phosphorus (C/N/P) atoms = 106/16/1). Although it is largely unknown whether DOM is produced in Redfield stoichiometry, our data show that DOM is depleted in phosphorus throughout the entire water column relative to Redfield values (Table 1). The dramatic increase in C/P and N/P ratios with depth indicates that phosphorus is preferentially regenerated from DOM, which implies that DOP must ultimately cycle more efficiently than either DOC or DON.

Selective removal of phosphorus from DOM presumably reflects the nutrient demands of marine microorganisms, suggesting that phosphorus is a limiting nutrient in surface waters of oligotrophic regions.

Lauren Lisa Clark, Ellery D. Ingall, Ronald Benner

University of Texas, Marine Science Institute, Port Aransas, Texas 78373, USA
email: ingall@utmsi.utexas.edu

- Redfield, A. C. *Am. Sci.* **46**, 205–222 (1958).
- Broecker, W. S. & Peng, T. S. *Tracers in the Sea* (Eldigio, Palisades, NY, 1982).
- Smith, S. V., Kimmerer, J. & Walsh, T. W. *Limnol. Oceanogr.* **31**, 161–167 (1986).
- Karl, D. M. & Yanagi, K. *Limnol. Oceanogr.* **42**, 1398–1405 (1997).
- Benner, R., Pakulski, J. D., McCarthy, M., Hedges, J. I. & Hatcher, P. G. *Science* **255**, 1561–1564 (1992).
- Benner, R., Biddanda, B., Black, B. & McCarthy, M. *Mar. Chem.* **57**, 243–263 (1997).
- McCarthy, M. D., Hedges, J. I. & Benner, R. *Chem. Geol.* **107**, 503–507 (1993).
- Amon, R. M. W. & Benner, R. *Nature* **369**, 549–552 (1994).
- Hori, T., Horiguchi, M. & Hayashi, A. *Biochemistry of Natural C–P Compounds* 1–200 (Maruzen, Shiga, 1984).
- Quin, L. D. in *Topics in Phosphorus Chemistry* (eds Grayson, M. & Griffith, E. J.) 23–48 (Wiley, New York, 1967).
- Kennedy, K. E. & Thompson, G. A. Jr *Science* **168**, 989–991 (1970).
- Rosenberg, H. in *Form and Function of Phospholipids* (eds Ansell, G. B., Hawthorne, J. N. & Dawson, R. M. C.) 333–344 (Elsevier, New York, 1973).
- Kittredge, J. S., Horiguchi, M. & Williams, P. M. *Comp. Biochem. Physiol.* **29**, 859–863 (1969).

American pie and food for thought

We Are What We Eat: Ethnic Food and the Making of Americans

by Donna R. Gabaccia

Harvard University Press: 1998. Pp. 278.

£16.50, \$24.95

Anne Murcott

The academic study of food and eating will have achieved a legitimate place among scholars when authors no longer need to defend their subject matter from accusations of triviality or frivolity, and when writers of dust-jacket blurbs stop having to describe a book on its social history as "thoroughly entertaining". Despite still having to suffer these minor indignities, Donna R. Gabaccia's book goes a long way towards making them obsolete.

Like others in the field, it considers the puzzle that human beings are simultaneously conservative and adventurous eaters. Like others, it plays about with Anthelme Brillat-Savarin's aphorism of 1825: "Tell me what you eat: I will tell you what you are." These beginnings are the springboard for its examination of about two hundred years of recurrent ethnic exchanges of both food and people that will reveal the identity of Americans. In the process, though, this book takes us further than others by exemplifying some intellectual attitudes that are well worth considering.

Organized largely chronologically, the book wisely avoids puns on melting-pots and cooking-pots, relying instead on creolized eating, a notion used here in a loose sense without reference to dominant or dominated groups. It begins in colonial times, takes us through the most fevered period of the mixing of cultures, tastes and immigrants (roughly from the 1880s to the Second World War), past a self-conscious 1970s reassertion of ethnic variety, to reach the present and the mildly surprising claim that the "foods we eat commemorate a long history of peaceful cultural interaction".

Plenty of thought-provoking and probably little-known details are presented along the way. For instance, slaves would sell their surplus home-grown vegetables and fresh meat to their masters. Both Chinese and non-Chinese lunch counters in a small area of Massachusetts served chow-mein sandwiches, an enterprising 1920s invention; as if it were not enough of a cross-cultural blend to mix chow-mein noodles, bean sprouts, chopped meat, onions and gravy in a hamburger bun, non-meat versions were available for Roman Catholic customers on a Friday. And later in this century, in an anonymous memoir, an Italian boy wrote that it never occurred to him that "just being a citizen of the United States meant that I was an



Pizza the action: contestants at the first world pizza-eating championships held in New York in 1996.

'American'. 'Americans' were people who ate peanut butter and jelly on mushy white bread that came out of a plastic package."

Gabaccia has a lightness of style, but this should not beguile readers into thinking that this is just a pleasing story-book with vivid illustrations. It is a skilfully written professional history imbued with a social anthropological sensibility. I wish that more British social anthropologists (and sociologists) in this field would trouble themselves to return the compliment by paying such diligent attention to social history. Gabaccia not only embraces the anthropological insight that human beings bestow meaning on food, making it not just good to eat but also good to communicate with, but goes on to grasp the other side of the anthropological debate, which requires detailed analysis of the material and economic circumstances that bring people and food together to allow communicative meanings to be created.

But more than this, Gabaccia recognizes that understanding eating habits requires not just one but several histories: of recurring human migrations, of agriculture, of (big) business and of consumption. This intellectual attitude and methodological grip on the study of food and eating is the book's great strength. It is this attitude — not its substantive, and slightly unconvincing, conclusions, including the declaration that Americans' eating habits show them to be a nation of multiethnics rather than a multiethnic nation — that guarantees the book's respectable academic scholarship.



Early burger: the first McDonald's hamburger stand opened in 1948 in California.

And it is this attitude that students would do very well to follow.

The book will help to give students a good start. But it is frustrating to find so lean an index, and that not all sources are referenced. Was this to try and keep the price

down, or anxiety lest a greater weight of footnotes put off a popular market? Even if the latter, courting a wide readership is not to be confused with promoting the public understanding of science. A decision to reference everything properly would at least have spared author and editor the embarrassing slip of the pen that attributes *The Jungle*, Upton Sinclair's 1906 fictional exposé of the conditions of the Chicago slaughterhouses, to his contemporary Sinclair Lewis.

But the book's appendix of sources is a fine entrée for beginners. The important references to the professional literature on the social history of eating in the United States will give novices all they need to start by themselves. Even better, for the specialist and potential graduate student alike, Gabaccia provides an enthusiastic couple of pages as an enticing guide to remarkable, apparently extensive but seemingly untapped archives. And anyone investigating industrialized eating could do a lot worse than bear in mind her shrewd observation about commercialized food consumption: "American eaters' search for the familiar and the novel became matters of consumer choice just as producers' and retailers' experiments with both innovation and traditional techniques became marketing strategies." □

Anne Murcott is in the School of Health and Social Care, South Bank University, 103 Borough Road, London SE1 0AA, UK.

Needing treatment like a hole in the head

Last Resort: Psychosurgery and the Limits of Medicine

by Jack D. Pressman

Cambridge University Press: 1998. Pp. 555. £40.00, \$49.95

Hugh Freeman

The history of psychiatry has been punctuated by the enthusiastic use of unpleasant, bizarre and dubiously helpful 'treatments', each of which has eventually been discarded. To those who question the whole legitimacy of psychiatry as a branch of medical practice, or who deny that mental illness really is an illness, it all seems to confirm their lowest opinions. In the twentieth century, replacing the revolving chairs and cold showers of previous eras, two treatment methods have figured most prominently in this demonology: electroconvulsive therapy (ECT) and, worst of all, leucotomy, also known as psychosurgery or lobotomy. To cut into a person's brain on rather dubious scientific grounds seems like the ultimate in medical imperialism.

But, as the late Jack Pressman

shows in this impressive but flawed work, the story is much more complex. The generally accepted view, which is over-simplified, is that the idea stemmed from the experimental work in the early 1930s of John Fulton, professor of physiology at Yale University in New Haven, Connecticut, who had been a pupil of Sir Charles Sherrington. Fulton showed that cutting nerve tracts under the frontal lobes of monkeys relieved 'experimental neuroses'. He reported these results at a conference in London in 1935, where the audience included Egas Moniz, who was not only an academic neurosurgeon (then a rare bird) but had also been Foreign Minister of Portugal. Moniz was inspired to apply the procedure to human subjects suffering from severe mental illness, and the first few cases suggested that this was a promising, perhaps revolutionary, method of treatment. He also invented X-ray examination of the cerebral circulation, and was eventually to be rewarded in 1949 with the Nobel prize for physiology or medicine.

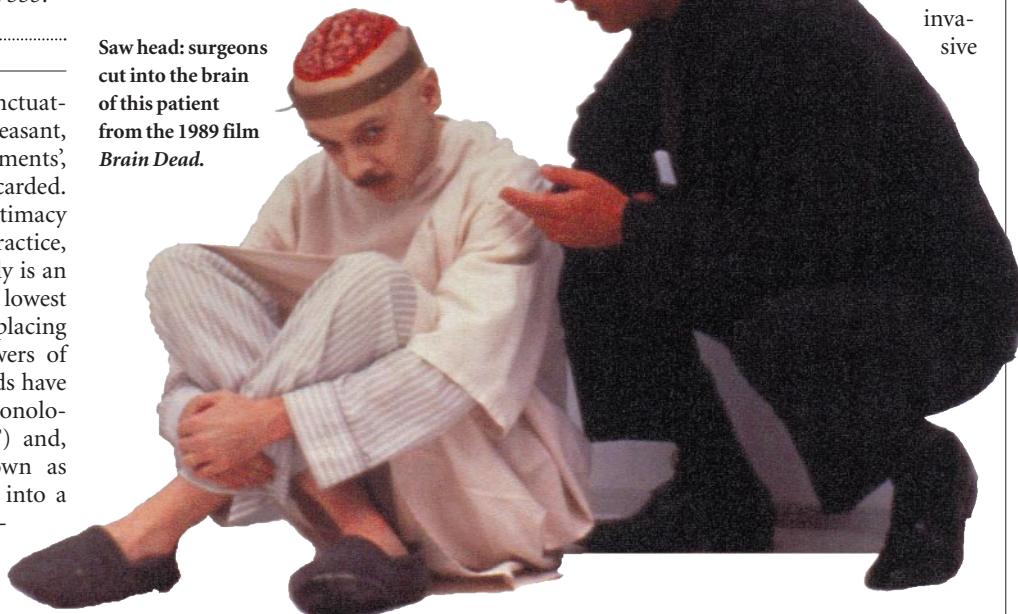
Pressman emphasizes that the introduction of leucotomy can be evaluated only against the actual circumstances of psychiatry at that time; it is easy with hindsight to be critical. The first specific treatment method in psychiatry — artificial malaria to treat tertiary syphilis — had been introduced only a few years before by Wagner von Jauregg in Vienna. This procedure was also unpleasant and uncertain in its effects. Insulin coma and convulsions induced by injected drugs were then in their early stages of development. Apart from these treatments there was nothing but sedatives (each with its own problems) and psychotherapy, which was ineffective as a treatment for severe mental illness. As a result, countless people suffered for years from distressing symptoms or disturbed behaviour for which nothing helpful could be done.

As with many innovations, the United States took up with enthusiasm what began elsewhere. Apart from the humanitarian need for effective new treatments, the state mental hospitals were full and, in the aftermath of the Depression, were desperately underfunded; anything that could make patients well enough for discharge would be of enormous value to them. Pressman suggests that there was a divergence between clinical and administrative aims, and this has been one of the many regular criticisms of leucotomy, but a patient who became well enough to leave the generally awful conditions of an overcrowded and understaffed state hospital was clearly benefiting as much as the institution.

The US pioneers of leucotomy were James Watts and Walter Freeman (no relation, but I met him towards the end of his career). Their early cases were mostly of agitated depression or obsessive-compulsive disorder, and the results were sometimes miraculous. But the operation became really widespread when it was extended to people with chronic schizophrenia, most of whom were in state hospitals. From early on there were warnings that the benefits were not gained without a price: intellect as such did not seem to be impaired, but there might be a blunting of the more sensitive aspects of the personality. Because of this, many variations were tried on the original 'standard' operation to try to relieve symptoms without causing such damage, with some success.

Leucotomy remained in vogue for about a decade in the United States, Britain and some other countries, but the discovery of major tranquillizers in the mid-1950s and the widespread use of electroconvulsive therapy largely removed the need for such an invasive

Saw head: surgeons cut into the brain of this patient from the 1989 film *Brain Dead*.



and potentially flawed procedure. Pressman's main point is that much of the condemnation of leucotomy has taken no account of its history, in that it ignores the clinical and administrative problems faced by those who used it and has an unreal view of the actual process of medical advance. This cannot always be on the basis of scientific deduction. But when Pressman finally asks "Did it work?", the issue is evaded.

From reading this work, one would conclude that the leucotomy story ended in about 1955, but this is not so. 'Modified' operations, with much more limited effects, continued to be used in Europe for intractable cases of depression and obsessive-compulsive disorder until they were largely replaced by more recent drugs. Sometimes they really did save lives. But Pressman makes hardly any reference to work outside the United States, giving only a passing mention to a British 1947 survey of 1,000 cases, which was at the time the largest and most systematic research carried out. He tells us how to read the story, but not how it ended, even though the text is long and repetitious.

Regrettably, Pressman died shortly after finishing this work. Had he lived, he would undoubtedly have made further important contributions to medical history. □

Hugh Freeman is at 21 Montague Square, London W1H 1RE, UK.

Genera obscura

The Platypus and the Mermaid: And Other Fingments of the Classifying Imagination

by Harriet Ritvo

Harvard University Press: 1997. Pp. 288.

£19.95, \$29.95

Peter J. Bowler

Harriet Ritvo's book about the Victorians' efforts to classify the world of animal life is fascinating but ultimately frustrating. This is not a history of scientific taxonomy as such, although it does deal in part with biologists' efforts to confront the unusual and the exotic. Human beings always seek to impose order on diversity, but who decides the categories to be imposed, and how do we deal with the exceptions that nature all too frequently provides? Scientific naturalists regarded it as their task to classify the animal kingdom, but they had to shift their ground in the face of newly discovered species that did not fit into existing schemes, to say nothing of hybrids and monstrosities.

In the course of Ritvo's story we meet eminent figures such as Charles Darwin and Richard Owen, but also a host of lesser scientific names now shrouded in obscurity. We also meet rival communities vying for the right to impose categories of their own,

including breeders, hunters, cooks and even railway companies. The cartoon used as a frontispiece shows a railway porter informing an old lady that cats, rabbits and parrots are counted as dogs, whereas her tortoise is an insect (and so, presumably, travels free).

Ritvo starts with the world of scientific taxonomy, pointing out the tensions even here between the rival methods of describing, naming and classifying species. The order that was supposed to have been imposed by the Linnean system turned out to be illusory, and in the early nineteenth century there were bitter debates between various schools of thought. But all the systems had to allow for the inclusion of exotic species that did not fit into the traditional categories, and it is here that we meet the platypus. An oviparous mammal was such a novelty that Owen at first refused to accept popular accounts that it laid eggs, and it was only in 1884 that W. H. Caldwell was able to confirm the stories and describe the eggs.

Ritvo breaks off at this point, paying no attention to the debates that raged among evolutionists trying to use both the monotremes and the marsupials as models for stages in mammalian phylogeny. Because 'missing links' were even more interesting if they were still alive, this area would have added another dimension to the story, and would have amplified the value-laden nature of the scientists' efforts to confront animals that seemed to inhabit the shadowy world between the major phyla.

A chapter on domesticated animals illustrates the tensions that could arise between the breeder and the scientist. The naturalists' binomial nomenclature was often ignored or even ridiculed by the practical men who dealt with horses, dogs and cattle every day. Some breeds achieved fame because they were thought to embody the original form of the wild ancestor from which domesticated varieties had arisen.

Hybrids were another popular topic, of great interest to Darwin. Ritvo recounts the complex lore of Victorian beliefs, observations and experiments on the question of how distinct races or even species could cross. This was not confined to animals: in an age acutely conscious of the alleged differences between the human races, the status of mulattoes and half-breeds was of perennial concern. Distinctive 'types' could be found (it was claimed) even among the inhabitants of the different English counties, although most of the population was, of course, by that time mixed.

Monstrosities, however defined, were a source of fascination to both scientists and the general public. Ritvo focuses mostly on attitudes towards, and ideas about, human deformities. Many unfortunates with gross



Jumping to conclusions...

bodily deformities were condemned to a life in shows and circuses, and, although a science of teratology emerged, it could do little to impose order on so bizarre a range of abnormal structures. Some individuals were seen almost as monsters even though they were merely unusually tall or fat. As a Leicester man, I was pleased to see the illustration of Daniel Lambert, perhaps the fattest man of his time, who evaded the worst experiences of the 'freak show' and was able to chat affably with his fashionable visitors.

It is in this chapter that the mermaid finally makes her (or rather its) appearance, because by this time the idea of a natural race of human-fish intermediates had been abandoned by all but the most vulgar. Specimens of mermaids were presented both to the public and to science, but were regarded as monstrosities, at least until they were exposed as frauds. Curiously, it is in this context that we meet a marine anomaly that was for a time a more serious bone of contention between scientists and other professionals. The sea serpent was sometimes portrayed as a relic of prehistoric times, but scientists such as Owen were anxious to discount it because they wanted to discredit the unsystematic observations of naval officers and the like.

Ritvo's book is a mine of fascinating information that will interest historians of both science and popular culture, although the former will sometimes be disappointed by its failure to follow through into a discussion of related issues, such as the roles of both intermediates and monstrosities in the emergence of evolutionary theory. More generally, the book's framework of analysis is less than satisfying because it is difficult to see a coherent pattern in the wealth of detail presented. The various discourses attempting to classify the variety of nature are themselves almost impossible to classify. □

Peter J. Bowler is in the Department of Social Anthropology at Queen's University, Belfast BT7 1NN, UK.

Learning to live with lasers

Lasers: Harnessing the Atom's Light

by James P. Harbison and Robert E. Nahory
W. H. Freeman: 1998. Pp. 214. \$34.95, £22.95

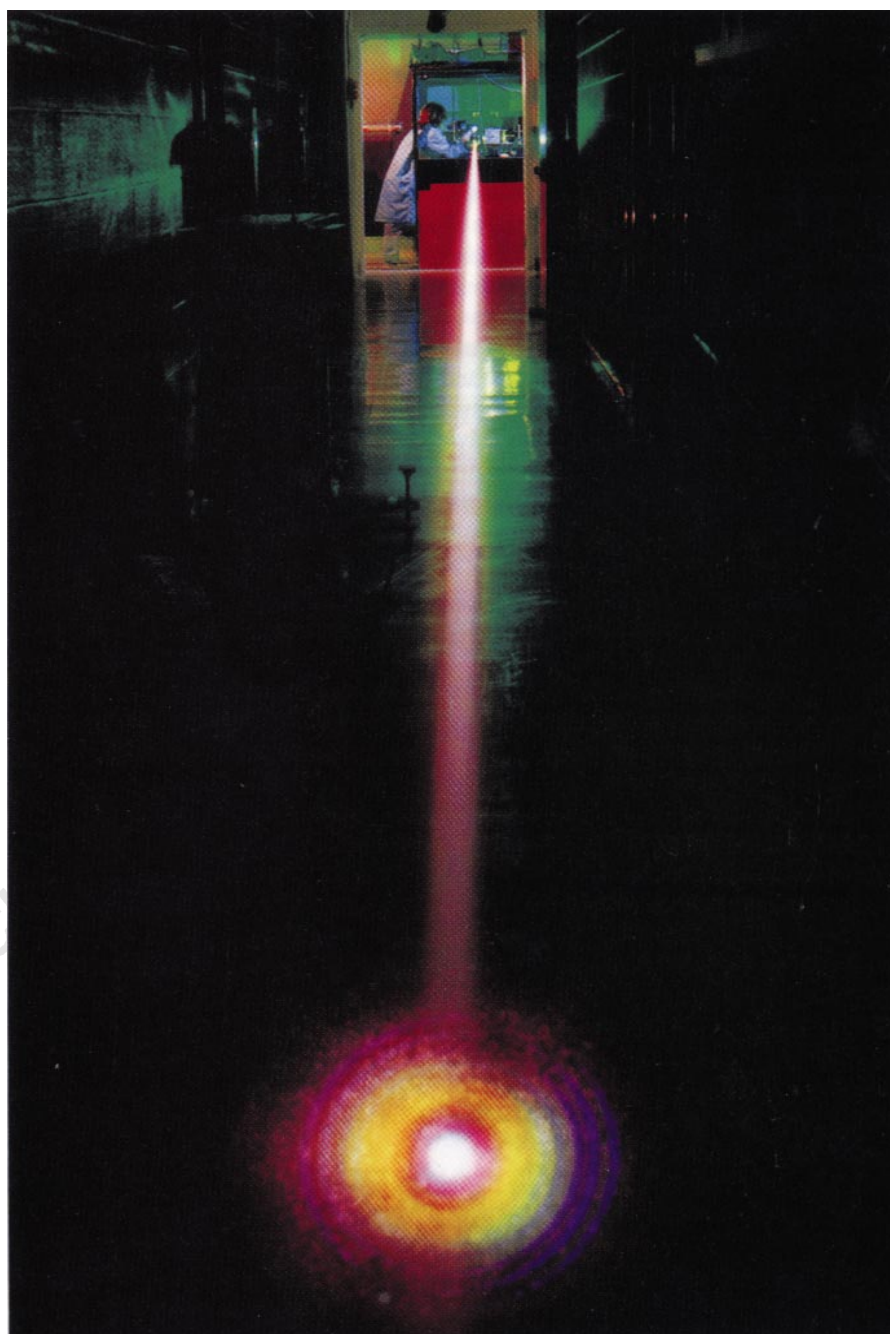
J. Christopher Whitehead

The laser was born in 1960 and one might think it should now be approaching middle age gracefully. But the pace at which new laser systems are being developed and new applications for lasers discovered suggests that it is undergoing a continual rebirth into a state of sophisticated maturity. Lasers are part of everybody's life yet we are often unaware of their existence, despite encountering them at the supermarket checkout and in the compact disc player, and relying on their power to bring high-speed communication into our homes.

James Harbison and Robert Nahory have produced an elegant book describing the many aspects of the laser, from its history and basic science to recent developments. Much of the book deals with the semiconductor lasers important in the field of high-speed communications. These lasers can be built to order by producing different layers of material only an atom or two thick. The true multidisciplinary nature of laser development is revealed as the authors describe the techniques of growing the crystals, the epitaxial methods of depositing the layers, and the processes of etching to remove selected parts of the device; this task involves scientists and engineers with skills in physics, chemistry, materials science and electrical engineering.

The trend is towards the microlaser, a million of which can be grown on a chip the size of a small fingernail. Quantum dot lasers have an active region consisting of a dot of indium gallium arsenide that is only 28 nanometres in diameter and 3 nanometres high. These devices are extremely compact, cheap to produce and use very little electrical power. At this scale, one can imagine such lasers being used in areas beyond communications, including optical computers and three-dimensional displays.

Harbison and Nahory also describe developments in lasers much larger in scale. These include X-ray lasers, which can be used to produce holograms of living cells by freezing an image in a billionth of a second; vacuum ultraviolet lasers, which can clean old paintings or 're-machine' the lens in a human eye to correct vision problems; lasers that emit short bursts of light lasting only a few femtoseconds, which can be used to investigate photosynthesis at a molecular level; and ultra-high-power lasers, in which a pulse from a bank of lasers is focused into a tiny pellet containing deuterium and tritium, creating pressures and temperatures akin to those



Tomorrow's world: a 40-metre plasma channel created by an ultrashort laser pulse, University of Michigan. Ultrashort pulse lengths could make tabletop X-ray lasers a reality.

at the centre of a massive star releasing energy in the process of laser fusion. They discuss the exciting ways in which lasers will aid understanding of many fundamental processes and provide solutions and opportunities that will transform the lives of many.

The authors understand the difficulties that non-specialists may have in understanding the concepts they discuss, and are imaginative in devising helpful analogies. They do this with the aid of first-class illustrations that make the book accessible to anyone with a background in science. □

J. Christopher Whitehead is in the Chemistry Department, University of Manchester, Manchester M13 9PL, UK.

New in paperback

In Search of the Big Bang: The Life and Death of the Universe

By John Gribbin

Penguin £8.99

Updated edition

Naked to the Bone: Medical Imaging in the Twentieth Century

By Bettyan Holtzman Kevles

Helix/Addison-Wesley, \$18

Reviewed by J. Mallard in *Nature* 386, 567 (1997).

A new surface electron-emission mechanism in diamond cathodes

M. W. Geis*, N. N. Efremow*, K. E. Krohn*, J. C. Twichell*, T. M. Lyszczarz*, R. Kalish†, J. A. Greer‡ & M. D. Tzabatzis*

* Lincoln Laboratory, Massachusetts Institute of Technology, Lexington, Massachusetts 02173-9108, USA

† Israel Institute of Technology, Technion City, Haifa 32000, Israel

‡ PVD Products, Bedford, Massachusetts 02173, USA

An electron-emission mechanism for cold cathodes is described based on the enhancement of electric fields at metal–diamond–vacuum triple junctions. Unlike conventional mechanisms, in which electrons tunnel from a metal or semiconductor directly into vacuum, the electrons here tunnel from a metal into diamond surface states, where they are accelerated to energies sufficient to be ejected into vacuum. Diamond cathodes designed to optimize this mechanism exhibit some of the lowest operational voltages achieved so far.

Conventional cathodes for applications from television to power transmitters use heat to boil electrons out of a metal into vacuum. However, these cathodes do not have the power efficiency or the dimensional stability to be used with micrometre-size structures, which are required for flat-panel displays and some power amplifiers. Cathodes that can be scaled to micrometre sizes use high electric fields instead of heat to pull electrons out of a solid into vacuum. The reliability and current density of these electric field emission cathodes depend upon both their geometry and the material used in their construction. Here we review field emission cathodes and show that a new cathode geometry which uses a novel material, diamond, has properties superior to those of previous cathodes.

For the cold cathodes we discuss, emission is obtained with a large electric field that causes electrons to tunnel over a potential barrier out of a metal substrate into vacuum. Material and fabrication techniques have both been used to increase emission by enhancing the electric field and reducing the barrier over which the electrons must tunnel. Excellent low electric field electron emission has been reported from diamond and amorphous diamond-like films on metal substrates, but practical application of these cathodes is limited by a serious lack of reproducibility^{1–3} and inconsistency (M. E. Kordes, personal communications). Emission originates from a few localized sites, which were believed to be due to the inconsistent bulk properties of the cathode material. The enhanced emission at the interface between the diamond surface, a conductive region, and vacuum, a new emission mechanism, may explain the localization of emission sites. If so, a discontinuous diamond film that provides an abundance of interfaces should be a better electron emitter than a continuous diamond film^{4,5}.

We now describe two generally accepted emission mechanisms: geometric electric-field enhancement⁶, and Schottky diode with a negative-electron-affinity semiconductor^{2,3}. Semiconductors and insulators have a negative electron affinity (NEA) if the minimum energy of electrons in the conduction band is above the minimum energy of electrons in vacuum. Experimental results are then described that cannot be explained by the previous emission mechanisms. A mechanism is proposed that combines the high electric fields that can be obtained at the intersection of a semiconductor surface, a metal substrate and vacuum (a so-called triple junction)^{7,8} with the mobile electron surface states that form on NEA surfaces^{9,10}. Cathodes fabricated to maximize the triple junction–NEA surface-emission mechanism, referred to here as surface emission cathodes, emit measurable currents (10^{-10} A cm⁻²) at gate voltages of 3 to 4 V, and useful current densities (>1 mA cm⁻²) at 6

to 10 V. In some cases, these electrons are emitted as a collimated beam of nearly monoenergetic electrons. With our present designs, some surface emission cathodes suffer from excessive gate current, which can vary from 0.2 to 10^5 times the emitted current.

Established electron-emission mechanisms

The most commonly used field emission technique is geometric electric field enhancement at a high aspect ratio conductive cone (Fig. 1a). In this approach, the field can be increased above 100 times the average applied electric field, making it possible to obtain emission with average applied fields of 1×10^3 to 1×10^6 V cm⁻¹, whereas theoretically no measurable emission would occur from a perfectly smooth surface at fields less than 1×10^7 V cm⁻¹. The emission current density, J , is related to the local electric field E at the emitting surface by the Fowler Nordheim equation

$$J = aE^2 e^{-b/E} \quad (1)$$

where a and b are constants used to fit the data. A theoretically more accurate equation that incorporates the work function of the metal has been derived⁶. However, accurate current calculations using this equation are nearly impossible, because the local geometrically

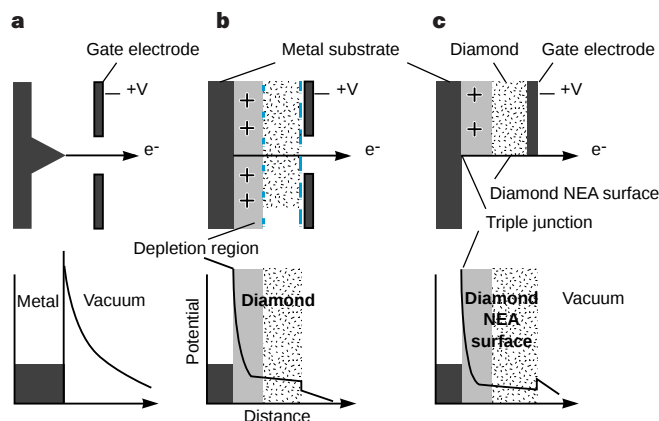


Figure 1 Diagrams and plots of potential electron energy versus distance for three electric field emission mechanisms. **a**, Geometric electric field enhancement; **b**, Schottky diode electric field enhancement; and **c**, surface emission cathode. Arrows marked 'e' indicate possible emission paths of electrons in the cathode structures.

enhanced electric field at the emitting surface is difficult to determine and the work function can vary significantly over an atomic scale. Polished stainless steel will emit measurable currents $> 1 \times 10^{-10}$ A, to an anode 1 mm in diameter spaced 100 μm away, at an applied field of $1 \times 10^6 \text{ V cm}^{-1}$. When the surface is roughened with 600-grit emery cloth, causing increased local field enhancement, the same currents are obtained at $\sim 1 \times 10^5 \text{ V cm}^{-1}$. As the maximum current that can be reliably obtained from a single geometrically enhanced field emission cathode site is $\sim 10 \mu\text{A}$, arrays of cathodes are generally used to obtain higher currents. Such arrays of metal cone cathodes with a metal grid structure spaced 100 nm to 10 μm apart require 20 to 200 V between the cones and the grid to obtain working current densities $> 1 \text{ mA cm}^{-2}$ (refs 11, 12).

A second mechanism (Fig. 1b) requires an electron-donor impurity-doped NEA semiconductor. This semiconductor forms a Schottky diode with the metal substrate. The emitted current is limited by tunnelling of electrons through the metal–semiconductor Schottky diode and not by electron emission from the semiconductor into vacuum. When an electric field is applied across the semiconductor, the dopants become positively ionized and form a depletion region at the metal–semiconductor junction. If type Ib diamond, which is typically nitrogen-doped to $2 \times 10^{19} \text{ cm}^{-3}$ (a deep donor)¹³, is used as this semiconductor, then for a bias voltage of 10 V, the depletion region is $\sim 20 \text{ nm}$ wide and the electric field at the metal–semiconductor interface is $\sim 1 \times 10^7 \text{ V cm}^{-1}$, which causes tunnelling of electrons the 1–5 nm distance²³ from the metal into the semiconductor. Like geometric electric-field enhancement, roughening the metal–semiconductor interface can further enhance emission, as discussed in detail elsewhere^{2,14–16}. The current density, J , is related to the semiconductor doping density, n , the barrier height at the metal–semiconductor interface Ψ , and the potential drop across the semiconductor, V , by

$$J \cong \frac{fnV^2}{\epsilon\Psi^2} \exp \left[-\frac{g\epsilon m_e^{1/2} \Psi^{3/2}}{V^{1/2} n^{1/2}} \right] \text{ for } V \gg \Psi \quad (2)$$

where m_e is the effective mass of the electron in the semiconductor, ϵ is the dielectric constant of the semiconductor, and f and g are constants used to fit the data. This equation is a simplified form of an equation¹⁷ derived using the WKB approximation⁶. The emission increases with the doping density and decreases with the dielectric constant, the effective mass of the electron, and the barrier height. Although the functional form of equation (2) differs from equation (1), both functions will usually fit the same emission data¹⁷, so it is difficult to differentiate between the two mechanisms from cathode characteristics. Once again, accurate emission current calculations are almost impossible, because too many of the parameters are not known to sufficient accuracy^{2,16}. Arrays of these cathodes emit current densities $\sim 1 \text{ mA cm}^{-2}$ when potentials of 10 to 20 V are applied between the metal substrate contacting the semiconductor and a metal gate structure spaced $\sim 1 \mu\text{m}$ away^{2,18}.

Experiments

A variety of experimental results discussed here have led us to consider a third emission mechanism which will be discussed later. One experiment used a type Ib (nitrogen doped to $\sim 1 \times 10^{19} \text{ cm}^{-3}$; ref. 13) cubo-octahedron diamond, faceted by (111) and (100) planes, $\sim 3 \text{ mm}$ on a side, which was implanted on the bottom, the (100) plane, with 34 keV Li^+ at 200 $^\circ\text{C}$ to a dose of $4 \times 10^{16} \text{ cm}^{-2}$ to enhance the electrical contact between the lower metal support and the diamond^{19,20}. The anode, which could be moved from touching to several centimetres above the diamond, consisted of either a molybdenum rod 0.5 mm in diameter or a square phosphor screen $\sim 1 \text{ cm}$ on a side. The diamond was prepared for emission by exposing it to an O_2 discharge, coating it with Cs, and re-exposing it to O_2 , a surface procedure known to enhance diamond's NEA

property, as discussed by P. E. Pehrsson (personal communication) and others^{21,22}. The anode was then positively biased until stable emission was obtained, which resulted in a yellow–green glow on the surface of the diamond, which originated from the ion-implanted region, and continued up the side of the diamond, around the top edge, and then across the top surface. Surface irregularities caused variations in the path of the glowing region. We believe this glow is caused by energetic electrons moving on the surface, some of which are scattered into the diamond, thereby exciting nitrogen, which then fluoresces. Once on the top surface, the electrons jump the 0–0.8 mm gap between the top diamond surface and the movable anode. When the anode was moved from touching the diamond to 0.5 mm above it, the anode voltage had to be increased by 12%, from 7 to 8 kV, to maintain a constant emission current of 1×10^{-5} . This indicates that most of the potential drop, $\sim 7 \text{ kV}$, appears across the diamond and is not being extended in the vacuum gap between the diamond and the anode²³. When a phosphor screen was used as an anode and placed on the diamond surface, the screen fluoresced where it met the diamond. This screen will fluoresce only when hit with electrons of energy $> 1 \text{ keV}$, implying that these electrons have considerable energy while still on the diamond surface. We propose that the electrons are field emitted from the implanted region into NEA semiconductor–vacuum surface states, as will be discussed later. Once in these states, the electrons can be accelerated to high energies, $> 1 \text{ keV}$. Electrons with similar energies appear to be emitted after travelling through the bulk of diamond and other semiconductors²³, but it is often difficult to differentiate electrons that move across the surface from those that are conducted through the bulk of the semiconductor.

To characterize the emission further, a thinner, 100- μm -thick, type Ib diamond was coated on the back side with Ni and graphitized on the top to form a conductive layer. The graphite layer was formed by sputtering the diamond surface with 1,200 eV Xe^+ . The diamond was then cleaved to obtain a clean, undamaged (111) oriented surface (Fig. 2, inset). After coating with Cs and exposing it with O_2 (ref. 21) a few kilovolts applied across the diamond between the Ni substrate and graphitized layer cause emission (Fig. 2). For this experiment, the graphite layer and the anode are at the same potential to minimize any possible field emission from the graphite layer. A phosphor screen anode placed above the diamond indicated that the electrons originated from the cleaved edge. As the electrons are emitted as a collimated beam, a retarding energy analyser was used to estimate the electron's energy, which is within 50 to 100 eV of the potential applied across the diamond, 1,000 to 6,000 V.

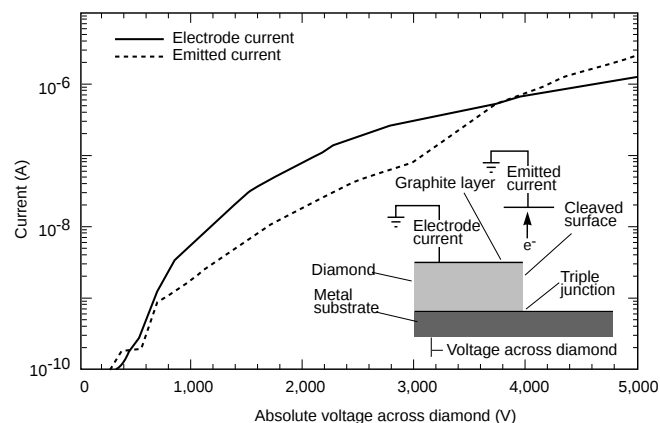


Figure 2 Plot of emitted and electrode currents as a function of voltage across the diamond for the structure shown in the inset. Inset, surface-emission cathode experiment using the cleaved surface of diamond to emit electrons.

As will be discussed, surface emission cathodes rely upon positive charges in the depletion region to create a large electron field at the triple junction (Fig. 1c and Fig. 3, inset). We have found from previous experiments that nitrogen impurities in diamond can be positively ionized and remain charged for days, but once exposed to light composed of photons of energy >2 eV (the ionization energy of the nitrogen impurity)¹³, photogenerated electrons in the diamond will neutralize these dopants. Figure 3 shows the effect of illumination; the emitted current is larger in the dark than in the light. In the dark, measurable emission occurs for potentials as low as 150 V across the diamond. With the light on, the emission is reduced and no measurable emission is obtained until the voltage across the diamond exceeds 1 kV. If the cathode is illuminated with no applied potential and then tested in the dark, the emission at the start of the experiment is nearly identical with the emission measured in the light. However, after the emission current exceeds 10^{-7} A, the behaviour changes and the characteristics are nearly identical with emission in the dark. This can be explained as follows. When the diamond is illuminated at low applied voltages, the dopants near the diamond surface are photoneutralized, so the electric field at the metal–diamond–vacuum triple junction is not sufficient to cause emission. At larger potentials, >2.5 kV, significant emission occurs, which can reionize these dopants faster than they can be photoneutralized. In this case, sufficient electric field will develop to enhance emission. Once the dopants are neutralized, emission even in the dark will be poor until sufficient potential is applied across the diamond to reionize them.

The effect of light on the emission from single-crystal diamond substrates as discussed above occurs in only about half of the diamonds tested. The cause of this inconsistency is unknown, but it may be the uncontrolled variation in the nitrogen doping of the type Ib diamonds used.

Surface emission cathodes

The new mechanism (Fig. 1c) is the composite of two known physical properties—the electric field enhancement at a triple junction^{7,8} and the high electron mobility at a NEA semiconductor surface–vacuum interface^{9,10}. Although emission for the two previous mechanisms can be related to device parameters by equations, no such equation exists for this new mechanism.

Triple junction. A triple junction is the intersection of a semiconductor surface with a metal substrate in vacuum. For this

application, an impurity-donor doped semiconductor is used in the triple junction geometry. When an electric field along the semiconductor surface is produced by a negative bias on the metal substrate, a substantial positive charge can form on the surface and in the bulk of the semiconductor near the triple junction (Fig. 1c). Part of this charge is the result of the Schottky diode that the semiconductor forms with the metal substrate, as previously discussed with the second mechanism (Fig. 1b). If the field is large enough, electrons will tunnel from the metal substrate onto the semiconducting surface with sufficient energy to cause secondary electron emission, which will further increase this positive charge. This model assumes n-type diamond, which does not exist. However, we have found that nitrogen functions as a deep electron donor, marking diamond n-type in character.

In many instances with other semiconductors, surface states pin the Fermi energy of the surface, independent of the semiconductor doping. If diamond has a pinned surface energy, then the depletion region may not extend to the diamond surface at the triple junction, contrary to the assumption in this model. These surface states are dependent on the chemistry on the semiconductor surface. It has been found that hydrogen²⁴ and oxygen²⁵ terminated diamond surfaces do not exhibit significant surface states, $<10^{12}$ cm⁻², to pin the Fermi energy. It is not known whether the oxygen–caesium terminated diamond surface has significant surface states that may pin surface energy.

When the field enhancement at the triple junction is combined with the high electron mobility on a NEA semiconductor surface^{9,10}, an emission mechanism can be constructed. This high mobility is the result of surface states, which will be discussed in more detail later, where electrons in vacuum are bound to a surface by their image charge in the bulk of the material, but are not allowed to pass through the surface because of quantum mechanical considerations. Emission from surface cathodes occurs when electrons tunnel from the metal substrate through the barrier at the triple junction. The height of the barrier is difficult to determine, but may be as little as 1 eV, or as much as the work function of the metal less the binding energy of the surface state. Thus the barrier height for tunnelling at the triple junction is less than that for tunnelling into vacuum. As will be discussed later, only 6 to 10 V across the surface is required to obtain currents equivalent to 1 mA cm⁻², whereas cathodes using the other mechanisms require 10 to 200 V to obtain the same currents.

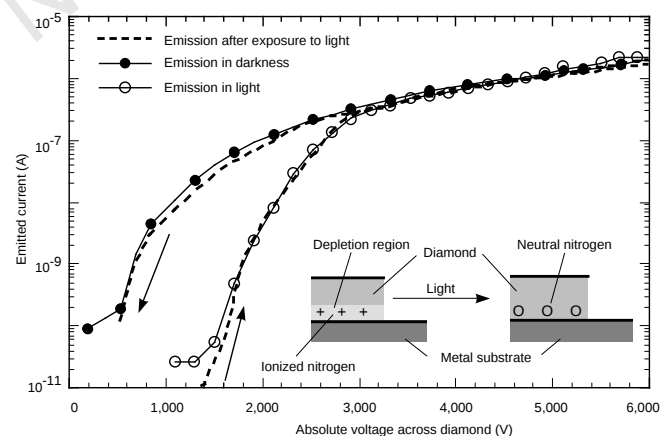


Figure 3 Emission current with and without ~ 10 mW cm⁻² of incandescent illumination from a fibre lamp. The dotted curve was obtained in the dark after 1 min of illumination. Arrows indicate the direction of voltage sweep for this curve. Inset, possible explanation of emission reduction with light, where illumination neutralizes positive charges near the triple junction.

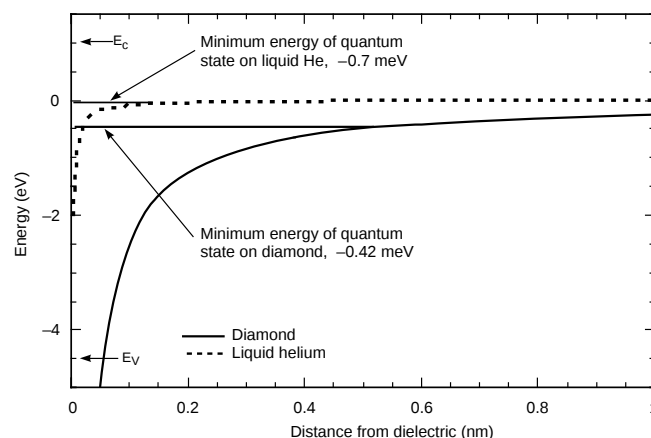


Figure 4 Potential energy curves as a function of distance from the liquid ⁴He and the diamond surfaces. The minimum energy states for an electron above liquid ⁴He and diamond surfaces are shown. E_c is the approximate minimum energy of electrons in the conduction band of liquid ⁴He and diamond. E_v is the maximum energy of electrons in diamond's valence band.

Surface states on NEA semiconductors. These surface states are a result of the NEA semiconductor–vacuum interface and have been studied since the 1930s (ref. 26). Electrons *in vacuo* are attracted to both dielectrics and metals by their image charge in the material. For NEA semiconductors, thermal electrons are not energetically able to pass into the semiconductor’s conduction band from the vacuum. Instead the electron resides in electrostatic surface states, 0.1 to 10 nm above the surface. These states were found on liquid ^4He , an NEA material, with the conduction band ~ 1 eV above the vacuum level on liquid ^3He , liquid Ne and metals^{27,28}. Figure 4 shows the electrostatic potential produced by the image charge for liquid ^4He and diamond. The surface states have binding energies, E in eV, approximated by

$$E = \frac{R(\epsilon - 1)^2}{16n^2(\epsilon + 1)^2}$$

where R is the Rydberg constant (13.7 eV), n is the quantum number, and ϵ is the dielectric constant of the NEA semiconductor (1.057 for liquid ^4He and 5.7 for diamond). The average distance, X , the electron is displaced from the interface when in the lowest energy state is given by

$$X = \frac{4a_0(\epsilon + 1)}{(\epsilon - 1)}$$

where a_0 is the Bohr radius, 0.0529 nm. For a liquid ^4He interface, the binding energy is 0.7 meV with the electron ~ 7.5 nm above the surface²⁹, and ~ 0.42 eV with the electron ~ 0.3 nm above the surface for diamond.

In this model, when electrons are accelerated on the semiconductor surface they may not remain bound on the diamond surface but move into the vacuum. However, this will have little effect on emission because the surface states are only required at the triple junction to keep the electrons out of the bulk of the diamond and to enhance tunnelling of the electrons from the metal substrate.

Practical cathodes

The previous experiments were all performed on comparatively thick diamond substrates. A more practical approach for low voltage emission is to use ion-beam assisted etching³⁰ to form

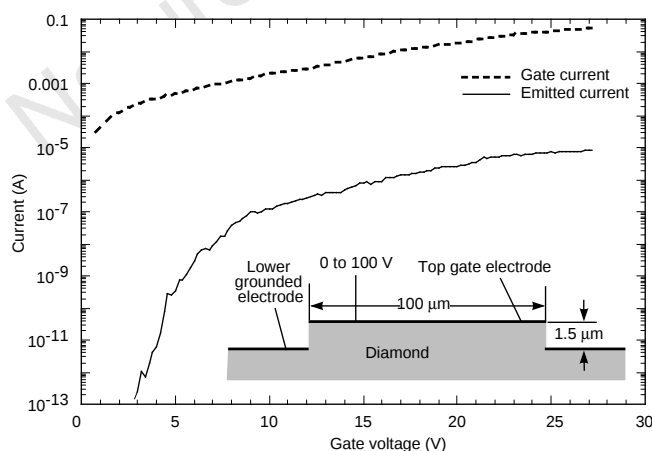


Figure 5 Emission and gate currents as a function of gate voltage for a surface emission device with a step height of 1.5 μm (inset). Current densities required for flat panel displays, $>1 \text{ mA cm}^{-2}$, $>10^{-7} \text{ A}$, are obtained at gate voltages of less than 10 V. Inset, cross-section of a diamond surface emitter. The emitter consists of type Ib diamond in which $100 \times 100 \mu\text{m}$ raised squares were formed by ion beam assisted etching. The structures are cleaned in molten NaNO_3 at 350°C and metallized by electron beam evaporation with 50 nm of Ni on the tops and bottoms, but the sides are nearly free of metal. The cathode was rinsed in a water solution of Cs_2CO_3 for low-voltage operation.

surface structures in diamond to obtain emission, as shown by the inset in Fig. 5. Instead of the 100- μm surface length used in the previous experiments (Figs 2, 3), 1.5–20- μm steps were etched into the diamond. The top and lower surface of the steps were coated with 50 nm of electron-beam-evaporated Ni. When a bias is applied across the step, measurable emission occurs at gate voltages <4 V and usable emission current occurs at voltages of 6 to 10 V. Light has no effect on the etched devices, but we are uncertain whether this is due to the difference in fabrication or the lower step heights. Both equations (1) and (2) can fit the emission current as a function of gate voltage¹⁷ and so the emission is Fowler Nordheim in character. Although these cathodes exhibit unexplained excessive gate currents from 100 to 10^5 times larger than the emitted current, the emission is far more consistent than previous diamond grit based cathodes². When the same cathode structures with 1- μm step heights are fabricated in thermally grown SiO_2 on a Si wafer instead of diamond, no measurable emission currents were obtained even at voltages exceeding 100 V, indicating that the choice of material used to separate the Ni electrodes is important. Further, surface emission cathodes made in type IIa (undoped) and type IIb (p-type boron doped) diamond, which contain fewer positively charged impurities at the triple junction than type Ib (nitrogen doped) diamond, require substantially higher voltages to obtain the same emission current.

Non-diamond surface emission cathodes

We next consider non-diamond surface emission cathodes as a background for the model used to explain diamond cathodes. Electron transport can occur on insulating surfaces and it has been suggested by K. Shoulders (personal communication) that this might be useful for cathodes. Surface emitting cathodes have been reported^{31,32} that operate by passing current through discontinuous thin films composed of metal islands, 1–1,000 nm in diameter³¹, on an insulating glass surface. Several materials can be used for this surface emitting cathode: metals (Au), semimetals (C)³³, and semiconductors ($\text{In}_x\text{Sn}_{1-x}\text{O}_2$)³⁴. Palladium oxide, PdO, surface cathodes have found use in flat panel displays³⁵.

Diamond and non-diamond surface emitting cathodes have much in common. Their geometries consist of metal electrodes separated by an insulating surface. Both cathodes exhibit the same electrical properties, such as an increase in emission with the addition of Cs or Cs salts³¹, emission of electrons with a narrow energy spread^{31,33,35}, low emission efficiency, and Fowler Nordheim emission characteristics. These similarities suggest that both types of cathodes operate with the same emission mechanism. Three models have been proposed: electron tunnelling into the vacuum at the enhanced electric field of the small (<10 nm diameter) metal islands³¹; thermal emission of hot electrons produced by electron conduction in the high electric fields between the metal islands³⁶; scattering and diffraction of energetic electrons out of the film by the metal islands³³.

Because the emission is related to the voltage across the film by the Fowler Nordheim equation (1), it has been argued³¹ that the emission mechanism is the result of electron tunnelling. However, Fowler Nordheim emission could also be the result of thermal emission³⁷. Thermal and tunnelling mechanisms could both play a role in emission³⁸, with thermal emission occurring primarily at voltages below 8 V, and electron tunnelling dominating at higher voltages. A large fraction of the electrons emitted into vacuum may be pulled back to the cathode by the local electric fields, where they could again be elastically scattered back into vacuum³⁹, which might explain the dependence of emission current on the anode voltage for PdO cathodes. None of these models addresses the narrow energy spread of the emitted electrons. Voltages across the films vary from 6 to 50 V and if the entire film emitted electrons, then the electron energy spread should be much larger than the reported ~ 2 eV (refs 32, 33, 37).

The model that we propose for diamond surface emission cathodes is an extension of the electron tunnelling model for surface film cathodes. In addition to the geometric electric field enhancement model identified earlier for non-diamond cathodes (Fig. 1a), a triple junction formed at the intersection of the doped semiconductor (diamond), metal and vacuum is used to model tunnelling of electrons from the metal contact to the semiconductor surface. For this tunnelling model, field enhancement is primarily caused by the positive space charge in the semiconductor at the triple junction. As discussed by Asai *et al.*³⁹, the electric field in the vacuum causes the electrons tunnelling into the vacuum to fall onto the cathode surface. However, instead of relying on elastic scattering to re-emit these electrons, we propose that a barrier exists at the diamond–vacuum interface which prevents the electrons from entering the diamond. The electrons are then accelerated along the diamond surface away from the triple junction, gaining sufficient energy to be emitted into vacuum.

Two other models might explain emission from the diamond cathodes. A surface discharge on the diamond could generate a plasma and hot electrons would be thermally emitted into the vacuum. The exceptionally high secondary electron yield of Cs-coated diamond^{22,40} could enhance electron generation, making a surface discharge more likely to form. If the discharge were responsible for emission, then the energy spread of the emitted electrons would approach that of the potential applied across the diamond. However, the energy spread is orders of magnitude smaller than the applied potential. Alternatively, emission could occur not from the triple junction, but from some asperity on the negative electrode, independent of the diamond substrate. In that case, the incandescent light would not affect emission, in contradiction to the experimental data shown in Fig. 3.

Discussion

An emission mechanism is discussed that relies on field emission of electrons at a triple junction onto the NEA surface of diamond. This emission mechanism requires the diamond surface to intersect the conducting substrate at a triple junction. It is interesting that Kordes⁴¹ in the first publication demonstrating low field emission from diamond said, “The area between the crystallites (which are easily identified as depressions or crevices ...) are strongly emitting areas.” That is, the emission does not occur from the tips of the diamond facets by geometric electric field enhancement, but from between the crystallites in the film where either the metal substrate or graphite layer separating the crystallites can serve as the conductive region of a triple junction. This suggests that many of the reported emission properties of diamond and diamond-like films are the result of surface emission from triple junctions and not the result of field emission from the bulk of the diamond into vacuum. Kumar has independently explained diamond and diamond-like field emission using a triple-junction model⁴.

The cathodes designed to maximize the new surface emission mechanism have more consistent emission than grit-based cathodes². These devices have some of the lowest reported operational voltages, can emit a nearly monoenergetic, collimated beam of electrons with energies up to 6 keV. The new mechanism explains the previously reported reduction of emission with light²¹. In addition to diamond, several other materials such as LiF (ref. 42), AlN (ref. 43), CaF₂ (ref. 44), BN (ref. 45), Cs-doped glass surfaces, and organic salts formed with alkali metals and crown ethers⁴⁶, may be useful for surface emission cathodes. □

Received 16 January; accepted 24 March 1998.

1. Kumar, N. Diamond Films '94 (abstr.) 5th Eur Conf. on Diamond, Diamond-like and Related Materials (Tuscany, Italy, September 25–30, 1994).
2. Geis, M. W., Twichell, J. C. & Lyszczyk, T. M. Fabrication and theory of diamond emitters. *Lincoln Lab. J.* **8**, 161–171 (1995).
3. Geis, M. W., Twichell, J. C. & Lyszczyk, T. M. Diamond emitters fabrication and theory. *J. Vac. Technol. B* **14**, 2060–2067 (1996).

4. Kumar, N. Diamond cold cathode technology for FED manufacturing (abstr.), DARPA High Definition Systems Information Exchange Conf. (Arlington, Virginia, March 23–26, 1997).
5. Geis, M. W., Efremow, N. N., Krohn, K. E., Twichell, J. C. & Lyszczyk, T. M. Diamond surface emission cathodes (abstr.), DARPA High Definition Systems Information Exchange Conf. (Arlington, Virginia, March 23–26, 1997).
6. Dyke, W. D. & Dolan, W. Field emission. *Adv. Electron. Electron Phys.* **8**, 89–185 (1956).
7. De Troureil, C. H. & Srivastava, K. D. Mechanism of surface charge of high-voltage insulators in vacuum. *IEEE Trans. Electr. Insul.* **EI-8**, 17–21 (1973).
8. Latham, R. V. (ed.) in *High Voltage Insulation: The Physical Basis* 229–240 (Academic, New York, 1981).
9. Cole, W. Properties of image-potential-induced surface states of insulators. *Phys. Rev. B* **2**, 4239–4252 (1970).
10. Grimes, C. C. Electron in surface states on liquid helium. *Surf. Sci.* **73**, 379–395 (1978).
11. Spindt, C. A., Brodie, L., Humphrey, L. & Westerberg, E. R. Physical properties of thin-film field emission cathodes with molybdenum cones. *J. Appl. Phys.* **47**, 5248–5263 (1976).
12. Bozler, C. O. *et al.* Arrays of gated field emitter cones having 0.32 μm tip-to-tip spacing. *J. Vac. Sci. Technol. B* **12**, 629–631 (1994).
13. Collins, A. T. & Lightowers, E. C. in *Properties of Diamond* (ed. Field, J. E.) 79–106 (Academic, New York, 1979).
14. Myers, A. F. *et al.* Characterization of amorphous carbon coated silicon field emitters. *J. Vac. Sci. Technol. B* **14**, 2024–2029 (1996).
15. Givargizov, E. I., Zhironov, V. V., Kuznetsov, A. V. & Plekhanov, P. S. Cold emission from the single-crystalline microparticle of diamond on a Si tip. *J. Vac. Sci. Technol. B* **14**, 2030–2033 (1996).
16. Lerner, P., Cutler, P. H. & Miskovsky, N. M. Theoretical analysis of field emission from a metal diamond cold cathode emitter. *J. Vac. Sci. Technol. B* **15**, 337–342 (1997).
17. Geis, M. W., Efremow, N. N., Krohn, K. E., Twichell, J. C. & Lyszczyk, T. M. Theory and experimental results of a new diamond surface-emission cathode. *Lincoln Lab. J.* **10**, 3–18 (1997).
18. Okano, K., Koizumi, S., Silva, S. R. P. & Amaratunga, G. A. J. Low-threshold cold cathodes made of nitrogen-doped chemical-vapour-deposited diamond. *Nature* **381**, 140–141 (1996).
19. Praver, S., Uzan-Spaguy, C., Braunstein, G. & Kalish, R. Can n-type doping of diamond be achieved by Li or Na ion implantation? *Appl. Phys. Lett.* **63**, 2502–2504 (1993).
20. Praver, S. & Kalish, R. Ion-beam-induced transformation of diamond. *Phys. Rev. B* **51**, 15711–15722 (1995).
21. Geis, M. W., Twichell, J. C., Macaulay, J. & Okano, K. Electron field emission from diamond and other carbon materials after H₂, O₂ and Cs treatment. *Appl. Phys. Lett.* **67**, 1328–1330 (1995).
22. Yater, J. E., Shih, A. & Abrams, R. Electron transport and emission properties of C(100). *Phys. Rev. B* **56**, R4410–R4413 (1997).
23. Geis, M. W., Twichell, J. C., Efremow, N. N., Krohn, K. & Lyszczyk, T. M. Comparison of electric field emission from nitrogen-doped, type Ib diamond, and boron-doped diamond. *Appl. Phys. Lett.* **68**, 2294–2296 (1996).
24. Pate, B. B. *et al.* The diamond (111) surface: A dilemma resolved. *Phys. B* **117/118**, 183–785 (1983).
25. Geis, M. W., Gregory, J. A. & Pate, B. B. Capacitance-voltage measurements on metal-SiO₂-diamond structures fabrication with (100)- and (111)-oriented substrates. *IEEE Trans. Electron Dev.* **38**, 619–626 (1991).
26. Davison, S. G. & Stejslicka, M. *Basic Theory of Surface States* (Clarendon, Oxford, 1992).
27. Osgood, R. M. & Wang, X. in *Solid State Physics* Vol. 51, 1–80 (Academic, New York, 1998).
28. McRae, E. G. Electronic surface resonances of crystals. *Rev. Mod. Phys.* **51**, 541–568 (1979).
29. Grimes, C. C. & Brown, T. R. Direct spectroscopic observation of electrons in image-potential states outside liquid helium. *Phys. Rev. Lett.* **32**, 280–283 (1974).
30. Efremow, N. N., Geis, M. W., Flanders, D. C., Lincoln, G. A. & Economou, N. P. Ion beam assisted etching of diamond. *J. Vac. Technol. B* **3**, 416–418 (1985).
31. Dittmer, G. Electrical conduction and electron emission of discontinuous thin films. *Thin Solid Films* **9**, 317–328 (1972).
32. Borziak, P. G., Kulyupin, Y. A., Nepijko, S. A. & Shamonya, V. G. Electrical conductivity and electrical emission from discontinuous metal films of homogeneous structure. *Thin Solid Films* **76**, 359–378 (1981).
33. Araki, H. & Hanawa, T. The temperature dependence of electron emission from a discontinuous carbon film device between silver film electrodes. *Thin Solid Films* **158**, 207–216 (1988).
34. Hartwell, M. & Fonstad, C. G. Strong electron emission from patterned tin–indium oxide thin films. *International Electron Devices Meeting* 519–521 (1975).
35. Yamaguchi, E. *et al.* Society for Information Display International Symposium Digest of Technical Papers **28**, 52–55 (1997).
36. Borziak, P., Kulyupin, Y. & Tomchuk, P. Electron processes in discontinuous metal films. *Thin Solid Films* **30**, 47–53 (1975).
37. Kulyupin, Y. A. & Nepijko, S. A. Effect of substrate deformation on electron emission from dispersed metal films. *Sov. Phys. Solid State* **17**, 1882–1884 (1976).
38. Blessing, R. & Pagnia, H. Electron emission from gold island films. *Phys. Status Solid B* **111**, 537–542 (1982).
39. Asai, A. *et al.* Society for Information Display International Symposium Digest of Technical Papers **28**, 127–130 (1997).
40. Mearini, G. T. *et al.* Stable secondary electron emission from chemical vapor deposited diamond films coated with alkali-halides. *Appl. Phys. Lett.* **66**, 242–244 (1995).
41. Wang, C., Garcia, A., Ingram, D. C., Lake, M. & Kordes, M. E. Cold field emission from CVD diamond film observed in emission electron microscopy. *Electron. Lett.* **27**, 1459–1460 (1991).
42. Lapiano-Smith, D. A., Eklund, E. A., Himpsel, F. J. & Teminello, L. J. Epitaxy of LiF on Ge(100). *Appl. Phys. Lett.* **50**, 2174–2176 (1991).
43. Benjamin, M. C., Wang, C., Davis, R. F. & Nemanich, R. J. Observation of negative electron affinity heteroepitaxial AlN on a(6H)-SiC(0001). *Appl. Phys. Lett.* **64**, 3288–3290 (1994).
44. Quiniou, B., Schwarz, W., Wu, Z., Osgood, R. M. & Yang, Q. Photoemission from thick overlying epitaxial layers of CaF₂ on Si(111). *Appl. Phys. Lett.* **60**, 183–185 (1992).
45. Powers, M. J. *et al.* Observation of a negative electron affinity for boron nitride. *Appl. Phys. Lett.* **67**, 3912–3914 (1995).
46. Dye, J. L. Electrides: Ionic salts with electrons as the anions. *Science* **18**, 272–273 (1990).

Acknowledgements. We thank K. Walters for suggesting the possibility of surface states on NEA semiconductors, R. Parker for suggesting the triple-junction mechanism for electron injection, H. Gray and J. Cuomo for discussion, and D. Lennon for technical assistance. This work was supported by the Ballistic Missile Defence Office through the Office of Naval Research and the Defense Advanced Research Projects Agency. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the United States Air Force.

Correspondence should be addressed to M.W.G. (e-mail: geis@LL.mit.edu).

Disruption of kilometre-sized asteroids by energetic collisions

E. Asphaug[†], S. J. Ostro[†], R. S. Hudson[‡], D. J. Scheeres[§] & W. Benz^{||}

^{*} Department of Earth Sciences, University of California, Santa Cruz, California 95064, USA

[†] NASA Jet Propulsion Laboratory 300-233, Pasadena, California 91109-8099, USA

[‡] School of Electrical Engineering and Computer Science, Washington State University, Pullman, Washington 99164-2752, USA

[§] Department of Aerospace Engineering and Engineering Mechanics, Iowa State University, Ames, Iowa 50011-3231, USA

^{||} Physikalisches Institut, Universität Bern, Sidlerstrasse 5, CH-3012 Bern, Switzerland

Recent numerical studies^{1–5} suggest that ‘rubble-pile’ asteroids (gravitationally bound aggregates of collisional debris) are common in the Solar System, and that self-gravitation may equal or exceed material cohesion for planetary bodies as small as several hundred metres. Because analytical scaling relations for impact cratering and disruption^{6–8} do not extend to this size regime, where gravity and material strength are both important, detailed simulations are needed to predict how small asteroids evolve through impact, and also to ascertain whether powerful explosions offer a viable defence against bodies headed for a collision with Earth. Here we present simulations, using a smooth-particle hydrodynamics code⁹, of energetic impacts into small planetary bodies with internal structure ranging from solid rock to porous aggregate. We find that the outcome of a collision is very sensitive to the configuration of pre-existing fractures and voids in the target. A porous asteroid (or one with deep regolith) damps the propagation of the shock wave from the impactor, sheltering the most distant regions, while greatly enhancing the local deposition of energy. Multiple-component asteroids (such as contact binaries) are also protected, because the shock wave cannot traverse the discontinuity between the components. We conclude that the first impact to significantly fragment an asteroid may determine its subsequent collisional evolution, and that internal structure will greatly influence attempts to disrupt or deflect an asteroid or comet headed towards Earth.

Target shape matters greatly during impacts¹⁰, so for the exploratory simulations presented here we adopt a representative asteroid instead of a sphere: specifically, a three-dimensional shape model reconstructed¹¹ from radar images of the near-Earth asteroid 4769 Castalia. This peanut-shaped asteroid is 1.6 km in longest dimension, and is one of several Earth-crossing objects imaged in detail by radar¹². Our three Castalia-shaped targets are made of (1) solid rock, (2) a pair of solid rocks separated by rubble, and (3) a 50% porous agglomeration of large boulders. These encompass several primary possible internal configurations of an asteroid. The material equation of state in each case is that of lunar gabbroic anorthosite^{13,14}, but substituting a specific density $\rho = 2.1 \text{ g cm}^{-3}$ for the solid targets and 4.2 g cm^{-2} for the porous target, for a constant target mass of $1.2 \times 10^{15} \text{ g}$. Elastic moduli and flaw-distribution parameters are derived from laboratory experiments in basalt^{15,16}. (We do not suggest that near-Earth asteroids are actually composed of basalt or anorthosite, but until adequate material descriptions become available for better analogues, we adopt this laboratory-verified ‘standard rock’ for simulations.) Our code, SPH3D (ref. 9), models shock propagation in elastic solids with a plastic yield criterion, and models explicit fracture and dynamic fragmentation under princi-

pal tension. The numerical resolution of each target is $\sim 130,000$ particles. Unless otherwise noted, each target is struck by an 8-m-radius, $5.8 \times 10^9 \text{ g}$ basalt sphere ($\rho = 2.7 \text{ g cm}^{-3}$) at 5 km s^{-1} . This speed is typical of asteroid collisions¹⁷ in the main belt, although for Earth-crossing objects a somewhat higher nominal impact speed is appropriate. These impacts are equivalent in energy to the 17 kiloton Hiroshima bomb. The mechanical efficiency of an impact is considerably greater than that of an explosion, however, and the effect of nuclear weapons on asteroids must be studied separately.

This impactor barely exceeds the disruption threshold for the non-porous intact target (Fig. 1). A $\sim 500\text{-m}$ -diameter damaged region forms, and distant fissures break the target into disconnected halves plus smaller pieces. We say that ‘disruption’ results when an intact asteroid is fragmented into pieces none more massive than half the original target, and that ‘dispersal’ results when less than half the original target’s mass remains gravitationally bound following the collision. A rubble-pile forms when disruption greatly exceeds dispersal. Only $\sim 10\%$ of the target mass exceeds the nominal $\sim 40 \text{ cm s}^{-1}$ escape velocity in this instance, so this event is not dispersive. For an irregular, rapidly-rotating asteroid like Castalia (period = 4 h), a detailed dynamical analysis is required to obtain the exact percentage. The impact imparts an impulse $\Delta v \sim 7 \text{ cm s}^{-1}$ to the non-escaping fraction’s centre of mass. Widespread surface fractures shown in Fig. 1 might become surface grooves such as those seen on the martian satellite Phobos and on

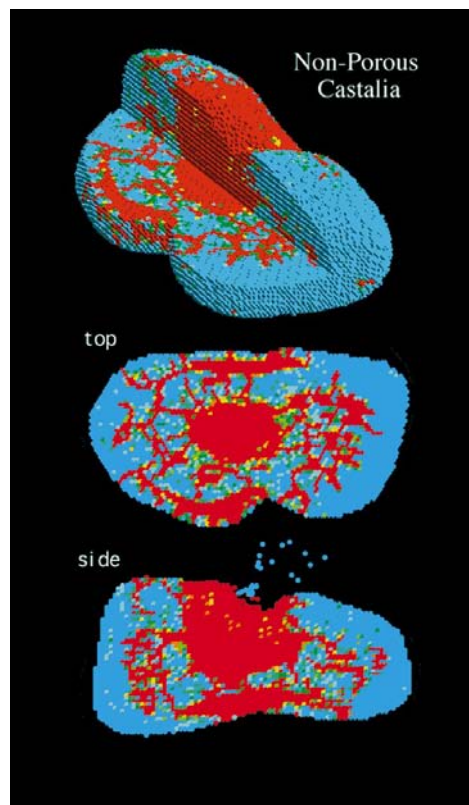


Figure 1 An initially intact Castalia (1.6 km longest dimension) seen 0.3 s after impact by an 8-m-radius basalt sphere at 5 km s^{-1} . The energy of this impact is equivalent to the Hiroshima explosion. Blue is unfractured rock; red is fully damaged rock, incapable of supporting tensile or shear stress. The shock wave has by this time crossed the asteroid twice, and has dissipated below the fracture threshold. Full hydrodynamic and dynamical evolution of the evolving crater bowl and the mobilized fragments would take hours of real time, and has not been modelled. Thin slices through the target centre are also shown: blue ejected particles in the side view are pieces of the impactor, and will escape. The non-escaping fraction ($\sim 90\%$) of the asteroid receives a Δv of $\sim 7 \text{ cm s}^{-1}$.

main-belt asteroids Gaspra and Ida^{18–21}. The diameter of the fully damaged region is the same size as the crater diameter predicted by gravity scaling²²—that is, if only gravity (and not strength) were a factor—implying that intact, rocky ~1-km bodies are near the size transition between strength- and gravity-dominance for large-scale cratering. Because strength has approximately equal influence here, ejecta velocities are greater than gravity scaling predicts, and ejecta deposited around this crater will be thin or absent.

The above simulation demonstrates that it is much easier to crack an intact ~1-km target in two than to disperse it. This could be an important mechanism for producing contact-binary forms among asteroids. Our next exploration is therefore the effect of an identical impactor striking one lobe of an initially binary Castalia, with intact lobes separated by ~20 m of fully damaged rubble (Fig. 2). In this case, due to impedance mismatch, shock energy is reflected from the discontinuity back into the struck lobe, which becomes utterly destroyed. The unstruck lobe suffers only minor damage. As before, ~10% of the target escapes; Δv is ~3 cm s⁻¹.

Porous targets can be similarly resistant to disruption, because shock wave propagation is hindered by irregularities and voids, resulting in localized energy deposition. However, a target which is porous and also strengthless might be comparatively easy to disperse. Figure 3 shows the aftermath of the same impactor striking the porous target, for comparison with Fig. 1; numerical resolution prevents us from separating the porous target into disconnected boulders, and hence the initial object is rigid. Almost all fracture damage (shown red) occurs within a ~500-m-diameter hemisphere centred on the impact point, with only minor damage far from the impact. More important differences are revealed by Fig. 4, which shows cross-sections of particle velocity and internal energy 1.2 s after projectile contact with the porous and non-porous targets. In the porous target, the shock dissipates rapidly with distance as work is done in the collapse of pores, and as scattering prevents coherent departure of shock energy from the impact zone. The resultant localized increase in particle speed and internal energy extends to a distance where the shock wave abruptly dies out. By comparison, the shock in the non-porous target (Fig. 1) is broadcast with few hindrances to the farthest reaches of the asteroid, leading to some disruption at great distance, and to lower energy deposition (kinetic

or thermal) near the impact. More than half the porous target is accelerated beyond nominal escape velocity, and a *strengthless* rubble pile would be just barely dispersed by this event, with a Δv of ~14 cm s⁻¹ applied to the remaining ~5 × 10¹⁴ g body. In our moderately-fragmented *rigid* porous body, however, only the damaged region can escape, which it does: no ejecta returns to fill the excavated crater or to form an ejecta blanket. As distal damage is hindered, a preponderance of large, pristine craters on an asteroid (such as 253 Mathilde²³) might imply a porous yet coupled interior, or else internal heterogeneity sufficiently great to scatter the shock and localize energy deposition.

The impact vapour and melt in our porous target penetrates the surface, rather than being directed immediately outwards. This produces an ejecta pattern which is far from conical (compare the trajectory vectors in Fig. 4); subsurface thermal alteration and material mixing is likely²⁴. A more precise equation of state, and a thermal conductivity model, is required in our code before quantitative predictions can be made about the degree of alteration and mixing, but the effect seems significant. A large meteoroid hitting a fine-grained porous target is less likely to result in such efficient vapour interpenetration, although laboratory experiments²⁵ show that energy deposition remains highly localized in these cases as well.

We conclude by examining scale-equivalence in impacts. Many

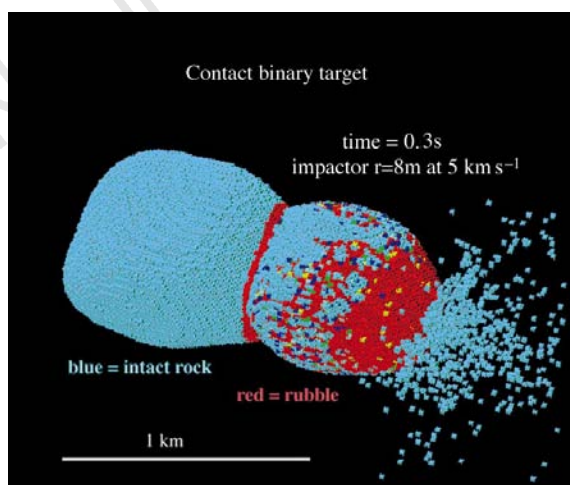


Figure 2 A contact-binary Castalia 0.3s after impact by an 8-m-radius basalt sphere striking at 5 km s⁻¹ on one end. The red band about the waist is pre-damaged, under-dense (1.7 g cm⁻³) material which presents an impedance barrier for the shock, thereby reflecting its energy back into the struck lobe. While damage to the struck lobe is almost total, very little damage occurs in the distal lobe. The non-escaping fraction (~90%) of the asteroid receives a Δv of ~3 cm s⁻¹. The final configuration will consist of an intact kernel surrounded by debris.

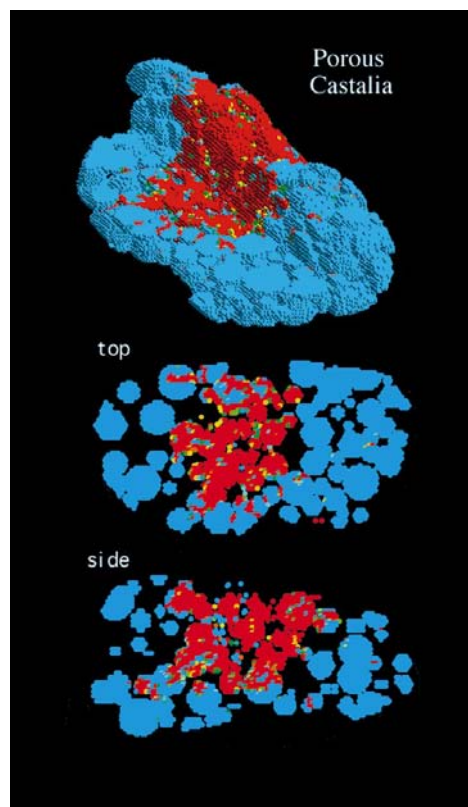


Figure 3 A 50% porous Castalia 0.3s after impact by an 8-m-radius sphere at 5 km s⁻¹. The target is rigid, formed from connected spheres with sizes ranging from ~30 to ~100 m. (Numerical resolution prevents us from adding a plane of rubble between each contacting sphere, in the manner of Fig. 2.) Top and side views are thin-sections, with black representing void space initially present in the porous configuration, not to be confused with impact disaggregation. Final fracture damage (red) is shown for comparison with Fig. 1, and is restricted to the region near the impact. More than half this target is accelerated beyond escape velocity, although only the fractured region can actually be mobilized. A true, disconnected rubble pile would be dispersed by this collision, with a ~14 cm s⁻¹ Δv applied to the ~5 × 10¹⁴ g non-escaping remainder.

aspects of a collision might be invariant to velocity provided that a 'coupling parameter'—essentially a hybrid between momentum (mv) and energy (mv^2)—is conserved⁷. Specifically, for solid rock targets, if $mv^{1.68}$ is held constant, the impact outcome should be the same. This relation has demonstrable merit in the case of hypervelocity cratering, at least in the gravity regime, provided that the impact speed is always much faster than the speed of sound in the target. However, our simulations show that for impact speeds much lower than this ($\sim 5 \text{ km s}^{-1}$ in our non-porous targets and $\sim 4 \text{ km s}^{-1}$

in our porous target), such invariances are poor descriptors of collisional outcome²⁶. Figure 5 compares the aftermath of a larger (20 m radius), slower (1 km s^{-1}) impactor, 'equivalent' to the 8 m, 5 km s^{-1} impactor of Fig. 1. In each case the asteroid cracks in two, but the predominant fractures for the slower projectile radiate from the impact, instead of propagating parallel beneath the surface. The greater amount of intermediate damage (shown yellow) and the smaller crater bowl (shown red) imply that impact disruption and cratering in the outer Solar System (where collision speeds are

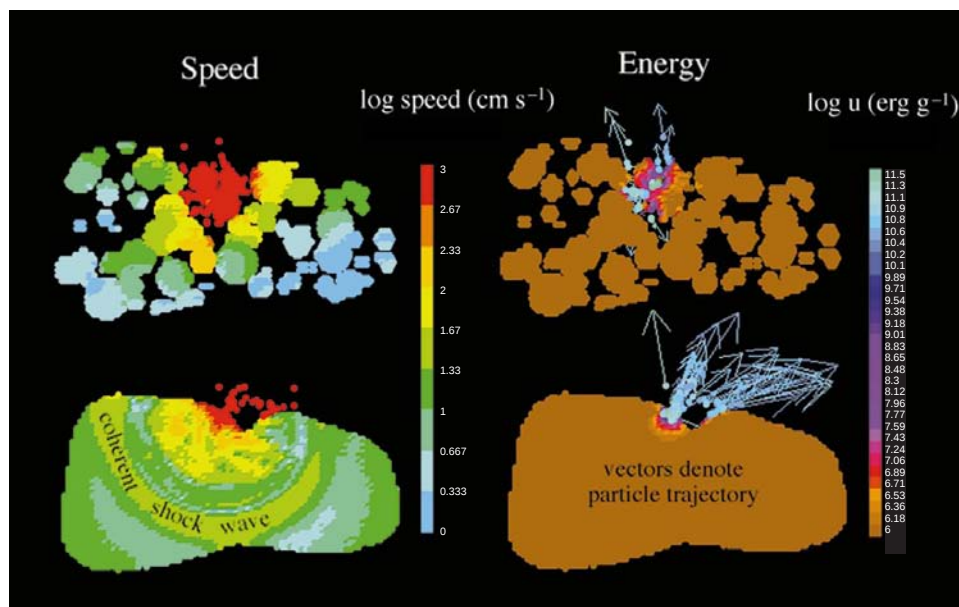


Figure 4 Effect of target porosity. Shown is a comparison of particle speed (left) and thermal energy (right) for the porous (top) and non-porous (bottom) targets. The target cross-sections are shown at $t = 0.12 \text{ s}$. At this time in the simulation, the shock wave has progressed as a relatively coherent signal to the far surface of the non-porous target, carrying much of the impact energy away from the contact zone and creating the spallation fractures seen in Fig. 1. In the porous target, by contrast, the shock is scattered and its energy confined by the voids, resulting in a

larger zone exceeding escape velocity (shown red and yellow) but lower intermediate velocities (shown green) and lower levels of distal disruption (Fig. 3). The entire damaged zone in the porous target exceeds escape velocity, and none of the far surface exceeds a few millimetres per second. The result will be a crater without an ejecta blanket, and minor seismic degradation in the distal regions of the target. Arrows in the energy plot indicate particle velocity, and show the dramatic effect of projectile interpenetration on ejecta trajectory.

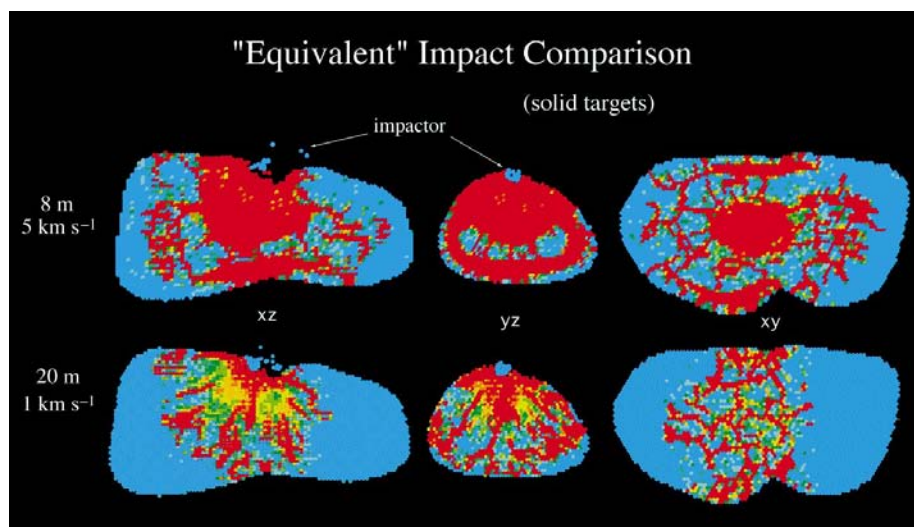


Figure 5 Effect of projectile size and speed. Shown is a comparison of fracture damage in the solid, non-porous Castalia for the original 8 m, 5 km s^{-1} projectile (top) versus a scale-equivalent 20 m, 1 km s^{-1} projectile striking at the same point. Although gross bulk aftermaths are the same (the target breaks in two, for instance), the surface-bounded spallation of the hypervelocity case (fractures

propagating sub-parallel to the surface) is replaced by radial fissuring in the subsonic impact (fractures radiating from the impact point), and the crater diameter in the subsonic impact is considerably smaller. Twice as much material (25%) is accelerated to escaping speed by the slower, larger impactor. The remainder which does not escape receives an impulse $\Delta v \sim 7 \text{ cm s}^{-1}$.

typically slower than $\sim 1 \text{ km s}^{-1}$) might differ significantly from what is assumed by current modelling efforts²⁷. The expected equation-of-state differences among small bodies (ice versus rock, for instance) presents another dimension of study; having recently adapted our code for massively parallel architectures (K. M. Olson and E.A., manuscript in preparation), we are now ready to perform a more comprehensive analysis.

The exploratory simulations presented here suggest that when a young, non-porous asteroid (if such exist) suffers extensive impact damage, the resulting fracture pattern largely defines the asteroid's response to future impacts. The stochastic nature of collisions implies that small asteroid interiors may be as diverse as their shapes and spin states. Detailed numerical simulations of impacts, using accurate shape models and rheologies, could shed light on how asteroid collisional response depends on internal configuration and shape, and hence on how planetesimals evolve. Detailed simulations are also required before one can predict the quantitative effects of nuclear explosions on Earth-crossing comets and asteroids, either for hazard mitigation²⁸ through disruption and deflection, or for resource exploitation²⁹. Such predictions would require detailed reconnaissance concerning the composition and internal structure of the targeted object. □

Received 4 February; accepted 18 March 1998.

- Asphaug, E. & Melosh, H. J. The Stickney impact of Phobos: A dynamical model. *Icarus* **101**, 144–164 (1993).
- Asphaug, E. *et al.* Mechanical and geological effects of impact cratering on Ida. *Icarus* **120**, 158–184 (1996).
- Nolan, M. C., Asphaug, E., Melosh, H. J. & Greenberg, R. Impact craters on asteroids: Does strength or gravity control their size? *Icarus* **124**, 359–371 (1996).
- Love, S. J. & Ahrens, T. J. Catastrophic impacts on gravity dominated asteroids. *Icarus* **124**, 141–155 (1996).
- Melosh, H. J. & Ryan, E. V. Asteroids: Shattered but not dispersed. *Icarus* **129**, 562–564 (1997).
- Housen, K. R., Schmidt, R. M. & Holsapple, K. A. Crater ejecta scaling laws: Fundamental forms based on dimensional analysis. *J. Geophys. Res.* **88**, 2485–2499 (1983).
- Holsapple, K. A. & Schmidt, R. M. Point source solutions and coupling parameters in cratering mechanics. *J. Geophys. Res.* **92**, 6350–6376 (1987).
- Housen, K. R. & Holsapple, K. A. On the fragmentation of asteroids and planetary satellites. *Icarus* **84**, 226–253 (1990).
- Benz, W. & Asphaug, E. Simulations of brittle solids using smooth particle hydrodynamics. *Comput. Phys. Commun.* **87**, 253–265 (1995).
- Asphaug, E. *et al.* Mechanical and geological effects of impact cratering on Ida. *Icarus* **120**, 158–184 (1996).
- Hudson, R. S. & Ostro, S. J. Shape of asteroid 4769 Castalia (1989 PB) from inversion of radar images. *Science* **263**, 940–943 (1994).
- Ostro, S. J. *et al.* Asteroid radar astrometry. *Astron. J.* **102**, 1490–1502 (1991).
- Ahrens, T. J. & O'Keefe, J. D. in *Impact and Explosion Cratering* (eds Roddy, D. J., Pepin, R. O. & Merrill, R. B.) 639–656 (Pergamon, New York, 1977).
- Tillotson, J. H. Metallic equations of state for hypervelocity impact. (General Atomic Report GA-3216, San Diego, 1962).
- Nakamura, A. & Fujiwara, A. Velocity distribution of fragments formed in a simulated collisional disruption. *Icarus* **92**, 132–146 (1991).
- Benz, W. & Asphaug, E. Simulations of brittle solids using smooth particle hydrodynamics. *Comput. Phys. Commun.* **87**, 253–265 (1995).
- Bottke, W. F., Nolan, M. C., Greenberg, R. & Kolvoord, R. A. Velocity distributions among colliding asteroids. *Icarus* **107**, 255–268 (1994).
- Belton, M. J. S. *et al.* Galileo encounter with 951 Gaspra—First pictures of an asteroid. *Science* **257**, 1647–1652 (1992).
- Belton, M. J. S. *et al.* Galileo's encounter with 243 Ida: An overview of the imaging experiment. *Icarus* **120**, 1–19 (1996).
- Asphaug, E. & Melosh, H. J. The Stickney impact of Phobos: A dynamical model. *Icarus* **101**, 144–164 (1993).
- Asphaug, E. *et al.* Mechanical and geological effects of impact cratering on Ida. *Icarus* **120**, 158–184 (1996).
- Housen, K. R., Schmidt, R. M. & Holsapple, K. A. Crater ejecta scaling laws: Fundamental forms based on dimensional analysis. *J. Geophys. Res.* **88**, 2485–2499 (1983).
- Veverka, J. *et al.* NEAR's flyby of 253 Mathilde: Images of a C asteroid. *Science* **278**, 2109–2112 (1997).
- Asphaug, E. *et al.* Impact evolution of icy regoliths. *Lunar Planet. Sci. Conf. (Abstr.) XXVIII*, 63–64 (1997).
- Love, S. G., Hörz, F. & Brownlee, D. E. Target porosity effects in impact cratering and collisional disruption. *Icarus* **105**, 216–224 (1993).
- Fujiwara, A., Cerroni, P., Davis, D. R., Ryan, E. V. & DiMartino, M. in *Asteroids II* (eds Binzel, R. P., Gehrels, T. & Matthews, A. S.) 240–265 (Univ. Arizona Press, Tucson, 1989).
- Davis, D. R. & Farinella, P. Collisional evolution of Edgeworth-Kuiper Belt objects. *Icarus* **125**, 50–60 (1997).
- Ahrens, T. J. & Harris, A. W. Deflection and fragmentation of near-Earth asteroids. *Nature* **360**, 429–433 (1992).
- Resources of Near-Earth Space* (eds Lewis, J. S., Matthews, M. S. & Guerrieri, M. L.) (Univ. Arizona Press, Tucson, 1993).

Acknowledgements. This work was supported by NASA's Planetary Geology and Geophysics Program.

Correspondence and requests for materials should be addressed to E.A. (e-mail: asphaug@earthsci.ucsc.edu).

Collective dynamics of 'small-world' networks

Duncan J. Watts* & Steven H. Strogatz

Department of Theoretical and Applied Mechanics, Kimball Hall,
Cornell University, Ithaca, New York 14853, USA

Networks of coupled dynamical systems have been used to model biological oscillators^{1–4}, Josephson junction arrays^{5,6}, excitable media⁷, neural networks^{8–10}, spatial games¹¹, genetic control networks¹² and many other self-organizing systems. Ordinarily, the connection topology is assumed to be either completely regular or completely random. But many biological, technological and social networks lie somewhere between these two extremes. Here we explore simple models of networks that can be tuned through this middle ground: regular networks 'rewired' to introduce increasing amounts of disorder. We find that these systems can be highly clustered, like regular lattices, yet have small characteristic path lengths, like random graphs. We call them 'small-world' networks, by analogy with the small-world phenomenon^{13,14} (popularly known as six degrees of separation¹⁵). The neural network of the worm *Caenorhabditis elegans*, the power grid of the western United States, and the collaboration graph of film actors are shown to be small-world networks. Models of dynamical systems with small-world coupling display enhanced signal-propagation speed, computational power, and synchronizability. In particular, infectious diseases spread more easily in small-world networks than in regular lattices.

To interpolate between regular and random networks, we consider the following random rewiring procedure (Fig. 1). Starting from a ring lattice with n vertices and k edges per vertex, we rewire each edge at random with probability p . This construction allows us to 'tune' the graph between regularity ($p = 0$) and disorder ($p = 1$), and thereby to probe the intermediate region $0 < p < 1$, about which little is known.

We quantify the structural properties of these graphs by their characteristic path length $L(p)$ and clustering coefficient $C(p)$, as defined in Fig. 2 legend. Here $L(p)$ measures the typical separation between two vertices in the graph (a global property), whereas $C(p)$ measures the cliquishness of a typical neighbourhood (a local property). The networks of interest to us have many vertices with sparse connections, but not so sparse that the graph is in danger of becoming disconnected. Specifically, we require $n \gg k \gg \ln(n) \gg 1$, where $k \gg \ln(n)$ guarantees that a random graph will be connected¹⁶. In this regime, we find that $L \sim n/2k \gg 1$ and $C \sim 3/4$ as $p \rightarrow 0$, while $L \approx L_{\text{random}} \sim \ln(n)/\ln(k)$ and $C \approx C_{\text{random}} \sim k/n \ll 1$ as $p \rightarrow 1$. Thus the regular lattice at $p = 0$ is a highly clustered, large world where L grows linearly with n , whereas the random network at $p = 1$ is a poorly clustered, small world where L grows only logarithmically with n . These limiting cases might lead one to suspect that large C is always associated with large L , and small C with small L .

On the contrary, Fig. 2 reveals that there is a broad interval of p over which $L(p)$ is almost as small as L_{random} yet $C(p) \gg C_{\text{random}}$. These small-world networks result from the immediate drop in $L(p)$ caused by the introduction of a few long-range edges. Such 'short cuts' connect vertices that would otherwise be much farther apart than L_{random} . For small p , each short cut has a highly nonlinear effect on L , contracting the distance not just between the pair of vertices that it connects, but between their immediate neighbourhoods, neighbourhoods of neighbourhoods and so on. By contrast, an edge

* Present address: Paul F. Lazarsfeld Center for the Social Sciences, Columbia University, 812 SIPA Building, 420 W 118 St, New York, New York 10027, USA.

removed from a clustered neighbourhood to make a short cut has, at most, a linear effect on C ; hence $C(p)$ remains practically unchanged for small p even though $L(p)$ drops rapidly. The important implication here is that at the local level (as reflected by $C(p)$), the transition to a small world is almost undetectable. To check the robustness of these results, we have tested many different types of initial regular graphs, as well as different algorithms for random rewiring, and all give qualitatively similar results. The only requirement is that the rewired edges must typically connect vertices that would otherwise be much farther apart than L_{random} .

The idealized construction above reveals the key role of short cuts. It suggests that the small-world phenomenon might be common in sparse networks with many vertices, as even a tiny fraction of short cuts would suffice. To test this idea, we have computed L and C for the collaboration graph of actors in feature films (generated from data available at <http://us.imdb.com>), the electrical power grid of the western United States, and the neural network of the nematode worm *C. elegans*¹⁷. All three graphs are of scientific interest. The graph of film actors is a surrogate for a social network¹⁸, with the advantage of being much more easily specified. It is also akin to the graph of mathematical collaborations centred, traditionally, on P. Erdős (partial data available at <http://www.acs.oakland.edu/~grossman/erdoshp.html>). The graph of the power grid is relevant to the efficiency and robustness of power networks¹⁹. And *C. elegans* is the sole example of a completely mapped neural network.

Table 1 shows that all three graphs are small-world networks. These examples were not hand-picked; they were chosen because of their inherent interest and because complete wiring diagrams were available. Thus the small-world phenomenon is not merely a curiosity of social networks^{13,14} nor an artefact of an idealized

model—it is probably generic for many large, sparse networks found in nature.

We now investigate the functional significance of small-world connectivity for dynamical systems. Our test case is a deliberately simplified model for the spread of an infectious disease. The population structure is modelled by the family of graphs described in Fig. 1. At time $t = 0$, a single infective individual is introduced into an otherwise healthy population. Infective individuals are removed permanently (by immunity or death) after a period of sickness that lasts one unit of dimensionless time. During this time, each infective individual can infect each of its healthy neighbours with probability r . On subsequent time steps, the disease spreads along the edges of the graph until it either infects the entire population, or it dies out, having infected some fraction of the population in the process.

Table 1 Empirical examples of small-world networks

	L_{actual}	L_{random}	C_{actual}	C_{random}
Film actors	3.65	2.99	0.79	0.00027
Power grid	18.7	12.4	0.080	0.005
<i>C. elegans</i>	2.65	2.25	0.28	0.05

Characteristic path length L and clustering coefficient C for three real networks, compared to random graphs with the same number of vertices (n) and average number of edges per vertex (k). (Actors: $n = 225,226$, $k = 61$. Power grid: $n = 4,941$, $k = 2.67$. *C. elegans*: $n = 282$, $k = 14$.) The graphs are defined as follows. Two actors are joined by an edge if they have acted in a film together. We restrict attention to the giant connected component¹⁶ of this graph, which includes ~90% of all actors listed in the Internet Movie Database (available at <http://us.imdb.com>), as of April 1997. For the power grid, vertices represent generators, transformers and substations, and edges represent high-voltage transmission lines between them. For *C. elegans*, an edge joins two neurons if they are connected by either a synapse or a gap junction. We treat all edges as undirected and unweighted, and all vertices as identical, recognizing that these are crude approximations. All three networks show the small-world phenomenon: $L \gg L_{\text{random}}$, but $C \gg C_{\text{random}}$.

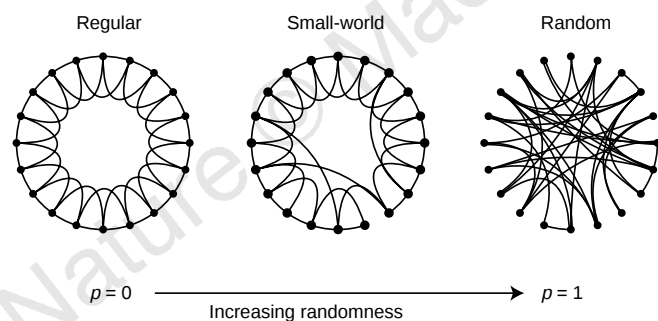


Figure 1 Random rewiring procedure for interpolating between a regular ring lattice and a random network, without altering the number of vertices or edges in the graph. We start with a ring of n vertices, each connected to its k nearest neighbours by undirected edges. (For clarity, $n = 20$ and $k = 4$ in the schematic examples shown here, but much larger n and k are used in the rest of this Letter.) We choose a vertex and the edge that connects it to its nearest neighbour in a clockwise sense. With probability p , we reconnect this edge to a vertex chosen uniformly at random over the entire ring, with duplicate edges forbidden; otherwise we leave the edge in place. We repeat this process by moving clockwise around the ring, considering each vertex in turn until one lap is completed. Next, we consider the edges that connect vertices to their second-nearest neighbours clockwise. As before, we randomly rewire each of these edges with probability p , and continue this process, circulating around the ring and proceeding outward to more distant neighbours after each lap, until each edge in the original lattice has been considered once. (As there are $nk/2$ edges in the entire graph, the rewiring process stops after $k/2$ laps.) Three realizations of this process are shown, for different values of p . For $p = 0$, the original ring is unchanged; as p increases, the graph becomes increasingly disordered until for $p = 1$, all edges are rewired randomly. One of our main results is that for intermediate values of p , the graph is a small-world network: highly clustered like a regular graph, yet with small characteristic path length, like a random graph. (See Fig. 2.)

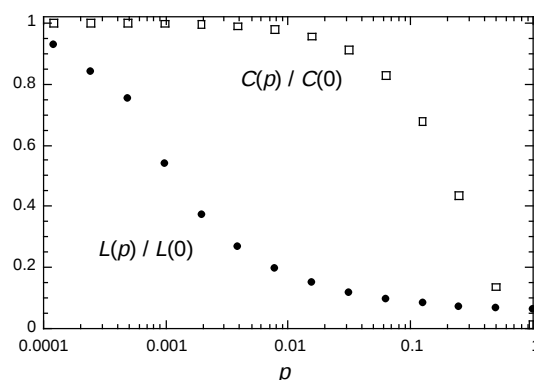


Figure 2 Characteristic path length $L(p)$ and clustering coefficient $C(p)$ for the family of randomly rewired graphs described in Fig. 1. Here L is defined as the number of edges in the shortest path between two vertices, averaged over all pairs of vertices. The clustering coefficient $C(p)$ is defined as follows. Suppose that a vertex v has k_v neighbours; then at most $k_v(k_v - 1)/2$ edges can exist between them (this occurs when every neighbour of v is connected to every other neighbour of v). Let C_v denote the fraction of these allowable edges that actually exist. Define C as the average of C_v over all v . For friendship networks, these statistics have intuitive meanings: L is the average number of friendships in the shortest chain connecting two people; C_v reflects the extent to which friends of v are also friends of each other; and thus C measures the cliquishness of a typical friendship circle. The data shown in the figure are averages over 20 random realizations of the rewiring process described in Fig. 1, and have been normalized by the values $L(0)$, $C(0)$ for a regular lattice. All the graphs have $n = 1,000$ vertices and an average degree of $k = 10$ edges per vertex. We note that a logarithmic horizontal scale has been used to resolve the rapid drop in $L(p)$, corresponding to the onset of the small-world phenomenon. During this drop, $C(p)$ remains almost constant at its value for the regular lattice, indicating that the transition to a small world is almost undetectable at the local level.

Two results emerge. First, the critical infectiousness r_{half} at which the disease infects half the population, decreases rapidly for small p (Fig. 3a). Second, for a disease that is sufficiently infectious to infect the entire population regardless of its structure, the time $T(p)$ required for global infection resembles the $L(p)$ curve (Fig. 3b). Thus, infectious diseases are predicted to spread much more easily and quickly in a small world; the alarming and less obvious point is how few short cuts are needed to make the world small.

Our model differs in some significant ways from other network models of disease spreading^{20–24}. All the models indicate that network structure influences the speed and extent of disease transmission, but our model illuminates the dynamics as an explicit function of structure (Fig. 3), rather than for a few particular topologies, such as random graphs, stars and chains^{20–23}. In the work closest to ours, Kretschmar and Morris²⁴ have shown that increases in the number of concurrent partnerships can significantly accelerate the propagation of a sexually-transmitted disease that spreads along the edges of a graph. All their graphs are disconnected because they fix the average number of partners per person at $k = 1$. An increase in the number of concurrent partnerships causes faster spreading by increasing the number of vertices in the graph's largest connected component. In contrast, all our graphs are connected; hence the predicted changes in the spreading dynamics are due to more subtle structural features than changes in connectedness. Moreover,

changes in the number of concurrent partners are obvious to an individual, whereas transitions leading to a smaller world are not.

We have also examined the effect of small-world connectivity on three other dynamical systems. In each case, the elements were coupled according to the family of graphs described in Fig. 1. (1) For cellular automata charged with the computational task of density classification²⁵, we find that a simple 'majority-rule' running on a small-world graph can outperform all known human and genetic algorithm-generated rules running on a ring lattice. (2) For the iterated, multi-player 'Prisoner's dilemma'¹¹ played on a graph, we find that as the fraction of short cuts increases, cooperation is less likely to emerge in a population of players using a generalized 'tit-for-tat'²⁶ strategy. The likelihood of cooperative strategies evolving out of an initial cooperative/non-cooperative mix also decreases with increasing p . (3) Small-world networks of coupled phase oscillators synchronize almost as readily as in the mean-field model², despite having orders of magnitude fewer edges. This result may be relevant to the observed synchronization of widely separated neurons in the visual cortex²⁷ if, as seems plausible, the brain has a small-world architecture.

We hope that our work will stimulate further studies of small-world networks. Their distinctive combination of high clustering with short characteristic path length cannot be captured by traditional approximations such as those based on regular lattices or random graphs. Although small-world architecture has not received much attention, we suggest that it will probably turn out to be widespread in biological, social and man-made systems, often with important dynamical consequences. □

Received 27 November 1997; accepted 6 April 1998.

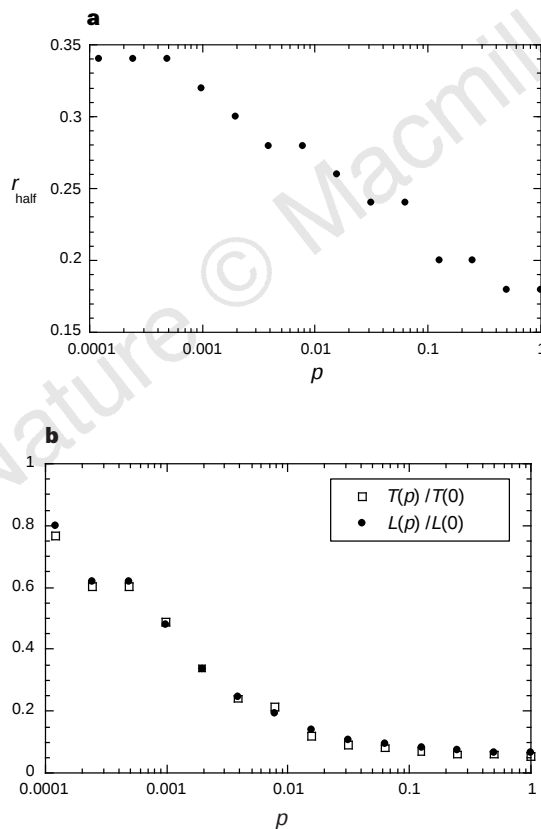


Figure 3 Simulation results for a simple model of disease spreading. The community structure is given by one realization of the family of randomly rewired graphs used in Fig. 1. **a**, Critical infectiousness r_{half} , at which the disease infects half the population, decreases with p . **b**, The time $T(p)$ required for a maximally infectious disease ($r = 1$) to spread throughout the entire population has essentially the same functional form as the characteristic path length $L(p)$. Even if only a few per cent of the edges in the original lattice are randomly rewired, the time to global infection is nearly as short as for a random graph.

1. Winfree, A. T. *The Geometry of Biological Time* (Springer, New York, 1980).
2. Kuramoto, Y. *Chemical Oscillations, Waves, and Turbulence* (Springer, Berlin, 1984).
3. Strogatz, S. H. & Stewart, I. Coupled oscillators and biological synchronization. *Sci. Am.* **269**(6), 102–109 (1993).
4. Bressloff, P. C., Coombes, S. & De Souza, B. Dynamics of a ring of pulse-coupled oscillators: a group theoretic approach. *Phys. Rev. Lett.* **79**, 2791–2794 (1997).
5. Braiman, Y., Lindner, J. F. & Ditto, W. L. Taming spatiotemporal chaos with disorder. *Nature* **378**, 465–467 (1995).
6. Wiesenfeld, K. New results on frequency-locking dynamics of disordered Josephson arrays. *Physica B* **222**, 315–319 (1996).
7. Gerhardt, M., Schuster, H. & Tyson, J. J. A cellular automaton model of excitable media including curvature and dispersion. *Science* **247**, 1563–1566 (1990).
8. Collins, J. J., Chow, C. C. & Imhoff, T. T. Stochastic resonance without tuning. *Nature* **376**, 236–238 (1995).
9. Hopfield, J. J. & Herz, A. V. M. Rapid local synchronization of action potentials: Toward computation with coupled integrate-and-fire neurons. *Proc. Natl Acad. Sci. USA* **92**, 6655–6662 (1995).
10. Abbott, L. F. & van Vreeswijk, C. Asynchronous states in neural networks of pulse-coupled oscillators. *Phys. Rev. E* **48**(2), 1483–1490 (1993).
11. Nowak, M. A. & May, R. M. Evolutionary games and spatial chaos. *Nature* **359**, 826–829 (1992).
12. Kauffman, S. A. Metabolic stability and epigenesis in randomly constructed genetic nets. *J. Theor. Biol.* **22**, 437–467 (1969).
13. Milgram, S. The small world problem. *Psychol. Today* **2**, 60–67 (1967).
14. Kochen, M. (ed.) *The Small World* (Ablex, Norwood, NJ, 1989).
15. Guare, J. *Six Degrees of Separation: A Play* (Vintage Books, New York, 1990).
16. Bollabás, B. *Random Graphs* (Academic, London, 1985).
17. Achacoso, T. B. & Yamamoto, W. S. *AY's Neuroanatomy of C. elegans for Computation* (CRC Press, Boca Raton, FL, 1992).
18. Wasserman, S. & Faust, K. *Social Network Analysis: Methods and Applications* (Cambridge Univ. Press, 1994).
19. Phadke, A. G. & Thorp, J. S. *Computer Relaying for Power Systems* (Wiley, New York, 1988).
20. Sattenspiel, L. & Simon, C. P. The spread and persistence of infectious diseases in structured populations. *Math. Biosci.* **90**, 341–366 (1988).
21. Longini, I. M. Jr A mathematical model for predicting the geographic spread of new infectious agents. *Math. Biosci.* **90**, 367–383 (1988).
22. Hess, G. Disease in metapopulation models: implications for conservation. *Ecology* **77**, 1617–1632 (1996).
23. Blythe, S. P., Castillo-Chavez, C. & Palmer, J. S. Toward a unified theory of sexual mixing and pair formation. *Math. Biosci.* **107**, 379–405 (1991).
24. Kretschmar, M. & Morris, M. Measures of concurrency in networks and the spread of infectious disease. *Math. Biosci.* **133**, 165–195 (1996).
25. Das, R., Mitchell, M. & Crutchfield, J. P. in *Parallel Problem Solving from Nature* (eds Davidov, Y., Schwefel, H.-P. & Männer, R.) 344–353 (Lecture Notes in Computer Science 866, Springer, Berlin, 1994).
26. Axelrod, R. *The Evolution of Cooperation* (Basic Books, New York, 1984).
27. Gray, C. M., König, P., Engel, A. K. & Singer, W. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature* **338**, 334–337 (1989).

Acknowledgements. We thank B. Tjaden for providing the film actor data, and J. Thorp and K. Bae for the Western States Power Grid data. This work was supported by the US National Science Foundation (Division of Mathematical Sciences).

Correspondence and requests for materials should be addressed to D.J.W. (e-mail: djw24@columbia.edu).

Size-controlled percolation pathways for electrical conduction in porous silicon

B. Hamilton*, J. Jacobs*, D. A. Hill†, R. F. Pettifer†, D. Teehan‡ & L. T. Canham§

* Department of Physics, UMIST, Manchester M60 1QD, UK

† Department of Physics, University of Warwick, Coventry CV4 7AL, UK

‡ CLRC Daresbury Laboratory, Daresbury, Warrington WA4 4AD, UK

§ DERA, Malvern, Worcester WR14 3PS, UK

Silicon shows photo- and electroluminescence at visible wavelengths when chemically etched into a microporous network of 'wires' several nanometres thick¹. This raises the possibility of a silicon-based optoelectronic technology. The luminescence properties may be understood on the basis of the injection or creation of electrons and holes in the interconnected network of wires which recombine radiatively with high efficiency^{1,2}. Elucidating the electron-transport mechanisms has been held back by several difficulties, particularly that of making stable, high-quality contacts to the porous material. Here we report experiments that probe the conduction process using photoemission stimulated by hard-ultraviolet/X-ray synchrotron radiation, obviating the need for good electrical contacts. We find that the conductivity of porous silicon films is temperature-dependent, and that the films become insulating at low temperatures. We suggest that these results may be understood in terms of a percolation process occurring through sites in the porous network in which conductivity is thermally activated, and we postulate that this activation may be the consequence of a Coulomb blockade effect^{3,4} in the

nanoscale channels of the film. This is consistent with our observation of optical 'unblocking' of conducting pathways. These results imply that the size distribution of the nanowires in the silicon backbone plays a key role in determining the conduction properties, and that porous-silicon light-emitting diodes may use only a small (and the least efficient) fraction of the material. Improvements in electroluminescence efficiency may be possible by taking into account the percolative nature of the conduction process.

These studies probe photoelectron emission in the hard-ultraviolet to X-ray (X-UV) region, with photon energies between 90 eV and 120 eV. Such energies probe the Si L₂₃ edge transitions. This ensures that the primary excitation probes material characterized by Si-Si bonds: that is, the silicon backbone of the porous film. The sensitivity of the L₂₃ edge to local chemistry has been noted previously^{5,6}. The X-ray attenuation length above the L edge for bulk silicon is ~500 Å, and the effects reported here were observed for films that are thicker than the attenuation depth, so that some fraction less than the whole thickness was directly excited by X-ray photons. We measured the total electron yield (TEY), which includes photoelectrons, primary Auger electrons, cascaded Auger electrons and secondary electrons. The TEY was measured by sensing the current flowing into the back of the Si wafer, which normally compensates exactly for the photoemitted electrons escaping from the surface. Such rapid charge compensation is always observed even for weakly conducting systems, because the small photoexcited current (typically 10⁻¹⁰ A) is easily supplied. But we have found that porous silicon films do not behave in this way; instead the TEY steadily decreases as temperature is reduced, and may fall below background noise levels at low temperatures.

TEY spectra obtained during a cooling cycle in which a conducting/insulating transition occurred are shown in Fig. 1. At the highest temperature (229 K), the usual features for such a measurement are observed⁷: the Si L₂₃ edge near 100 eV, together with some small

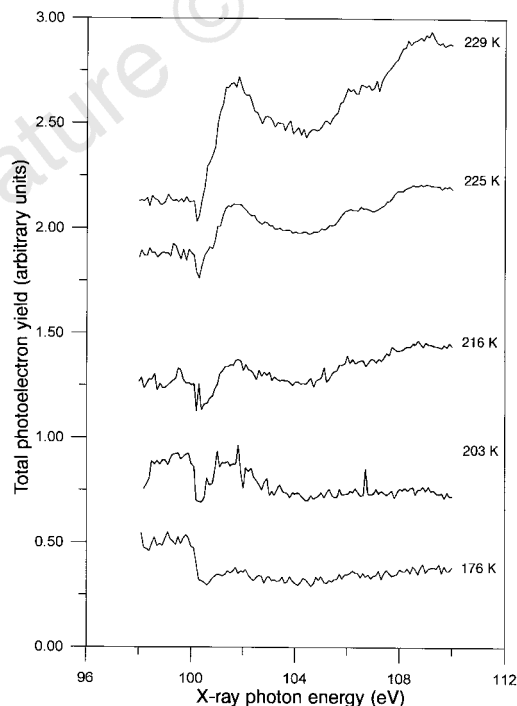


Figure 1 A typical example of the evolution of the Si L edge TEY spectra for an 80% porous sample as the temperature is lowered. The rapid loss of signal is due to the porous film entering a highly insulating state.

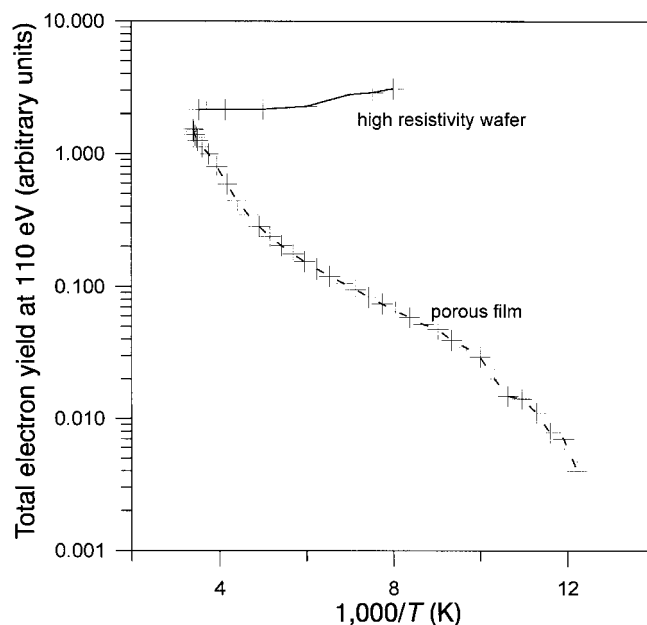


Figure 2 Temperature variation of the TEY signal for a 60% porous layer. Below 100 K, electron emission levels are below the measurement noise, indicating an absence of conducting pathways between the substrate and surface. The smooth loss of signal with reducing temperature is consistent with a gradual reduction of conduction pathways, indicative of percolation processes. Even at room temperature, the numbers of such pathways are still growing as the TEY is still increasing. Single-crystal silicon shows the opposite temperature dependence. Crosses are data points; broken line is a fit to data.

associated features. By 176 K, the L edge has disappeared, apart from an inversion at the threshold energy for L-edge emission. This inversion represents the system response to surface charging, associated with regions of the film that have become insulating. At the lowest temperature recorded in Fig. 1, this is the whole of the excited area.

The temperature variation of film conduction was obtained by measuring the TEY at fixed X-ray energy (110 eV) as a sample was heated. The result is shown in Fig. 2, for a sample with a porosity of ~60%. The data show that the integrated conduction increases smoothly with temperature from 100 K. Such behaviour is radically different to that observed for conventional silicon crystals, which is also shown in Fig. 2 for a high-resistivity wafer. The TEY for the Si wafer increases as the temperature is lowered (the opposite behaviour to that of the porous films), probably resulting from an increase in escape probability for electrons at lower temperatures. Clearly, within the temperature range measured, bulk silicon wafers are able to supply the small currents required to balance electron emission.

This gradual reduction in the ability of the porous films to supply electrons to the surface as temperature is lowered must reflect a change in electron transport within the film. It is well established that the conductivity of microporous material is orders of magnitude lower than that of bulk silicon⁸. Several models have been proposed to explain the conductivity of porous silicon, taking into account bandgap variation in the disordered quantum wires⁹, quantum size effects that lower the local dielectric constant (increasing dopant binding energies)¹⁰, and the existence of deep surface traps¹¹. Transient conduction¹² has also been measured and the time response has been partly explained in terms of percolation phenomena, which account for many properties of porous and granular systems¹³.

Our results indicate that percolation phenomena could underpin the gross features of conduction: the films become insulating when there is no longer a percolating conducting pathway through the

porous medium. This can be understood from the unusual way in which electron transport within porous films is sensed by X-ray excitation, shown schematically in Fig. 3. Here the porous film is developed into a network of sites that are 'blocked' and 'unblocked' with respect to electron transport. The unblocked sites form clusters within the network; clusters that span the film provide a percolating conducting pathway. The sample is uniformly illuminated with X-ray photons, but only the flux intersecting the surfaces of percolating clusters gives a TEY contribution. Three typical clusters are shown in Fig. 3: cluster A contains a small number of conducting sites and would present a high resistance, whereas the larger cluster B would be characterized by a much lower resistance. Cluster C is non-percolating and does not contribute to the TEY. Both percolating clusters have been shown to have the same active surface area for photoemission, and this allows us to understand why photoelectron-emission measurements are more sensitive to percolating geometries than contact-based measurements. The electron current drawn through a percolating cluster makes a TEY contribution that is controlled by the X-ray flux incident on its surface area but, crucially, the current drawn through a cluster will not depend on the size of that cluster as it threads down to the substrate. In contrast, contact-based measurements of conductivity are sensitive to the sizes of percolating clusters, as size controls the local resistance contribution. Clearly, the TEY measurement probes the material in a totally different manner to conventional conductivity experiments.

What is the mechanism that causes switching from 'unblocked' to 'blocked' sites as the temperature is lowered? Electrons passing through the structure find stable locations in the size fluctuations. In a single-electron picture, regions with a lower energy gap, which correspond to wider 'wires', form low-energy sites. Transferring an electron into a localized volume requires an energy $E = e^2/2C$, to be supplied in order to overcome the Coulomb barrier (here e is the electronic charge and C the self-capacitance of the 'isolated' volume). For nanometre-scale structures, the capacitance term is

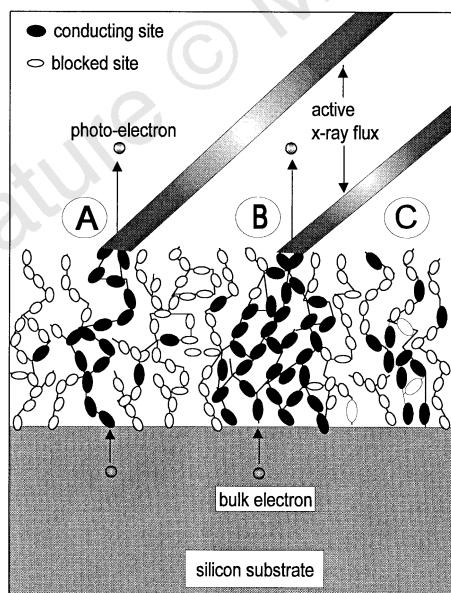


Figure 3 Diagram showing how photoelectron emission is sensitive to the fraction of the free surface connected to conducting routes or percolating clusters. Photoelectron emission is not sensitive to the internal volume of clusters; that is, the measurement does not sense conductivity. The sample is uniformly excited by synchrotron radiation, but only the flux intersecting conducting or fully percolating clusters is useful in producing electron emission. The scale of the Si backbone is greatly enlarged; in a real structure there would be a very larger number of percolation sites spanning the depth of the film. A, B and C represent three types of cluster described in the text.

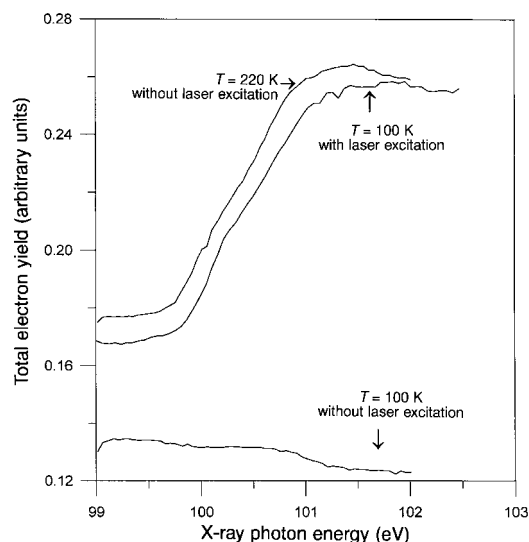


Figure 4 The effect of optically pumping porous layers which are in the insulating state (low temperature). The TEY spectrum is restored to its high-temperature form. This can be explained by photodischarge of nanostructures which have entered the Coulomb blockade regime and which renders the film locally insulating.

small, making E large. It is well established experimentally that Coulomb-blockade effects can dominate electron transfer through small structures^{3,4,14}. For nanoscale-structures with capacitance of ~ 1 aF (that is, 10^{-18} F), the charging energy becomes comparable with kT at 300 K, allowing observation of single-electron-transfer events at room temperature⁴. So Coulomb-blockade effects might be expected to exert a considerable influence on the transfer of electrons between the nanoscale wires in microporous silicon. As the temperature is lowered, the available thermal energy will become insufficient to allow electron transfer into sites possessing too large a Coulomb gap, unless the full voltage can be supplied locally. Thus the smaller (lower C) sites will switch off preferentially as the temperature drops, and a previously percolating cluster will switch into a non-percolating configuration. The skeleton size distribution measured for red-emitting p^+ porous silicon is broad, spanning 2 to 12 nm (ref. 15) which is consistent with Coulomb blockade effects occurring between temperatures of 100 K and 300 K.

In support of this picture, we find that films at low temperature in the insulating state can be routinely switched into the fully conducting state when simultaneously excited with visible laser light. This is shown in Fig. 4, which demonstrates complete restoration of the L edge. The films used to obtain the data for Figs 1 and 4 were similar: both were processed from p^+ material, were $\sim 80\%$ porous and were $\sim 8\mu\text{m}$ thick. The optical source used to switch between insulating and conducting states at low temperature was a 2 mW unfocused HeNe laser emitting at 631 nm. The optical absorption data¹⁶ for films of this type predict an optical extinction depth of several micrometres at this wavelength, ensuring that the optical radiation is absorbed throughout the film. This is consistent with the idea that processes occurring within the film, rather than on the surface, control the conducting-to-insulating transition. Optical unblocking of the conduction process is expected if a Coulomb blockade drives the switching mechanism. In the blocked state, individual sites must acquire a net negative charge. Although transfer of another electron into the site is forbidden, capture of a locally photoexcited hole is energetically favourable. The hole component of an electron-hole pair generated close to a blocked site may easily pass onto the site, reducing its local charge and therefore unblocking it. The electron from the original pair will move in the opposite direction to the hole in the local field and aids the conduction process. We note that optical perturbation of photoelectron emission has been reported previously^{17,18}; the mechanism suggested here is similar but operates at localized potential barriers rather than surface depletion fields.

The evidence for percolation effects in the transport of electron in porous silicon is important for the optimization of electroluminescent devices, which use conducting surface contacts. With such contacts, much of the current is likely to be shunted through the most massive percolating (lowest resistance) clusters, and it is to be expected that at room temperature these would be rich in the larger silicon sites offering the lowest Coulomb barriers. As the smallest nanoscale structures are known to be the most important for visible light emission, it is likely that existing contact technology is far from optimized. An improved light-emitting diode would require high uniformity in nanoscale structure size distribution. $\square$

Received 31 October 1997; accepted 2 March 1998.

- Canham, L. T. Silicon quantum wire array fabrication by electrochemical dissolution of wafers. *Appl. Phys. Lett.* **57**, 1046–1048 (1990).
- Hamilton, B. Porous silicon. *Semicond. Sci. Technol.* **10**, 1187–1207 (1995).
- Lee, P. A. Few electron nanostructures—a laboratory for studying strongly interacting systems. *Physica B* **189**, 1–5 (1993).
- Ali, D. & Ahmed, H. Coulomb blockade in a silicon tunnel junction device. *Appl. Phys. Lett.* **64**, 2119–2120 (1994).
- Sham, T. K. *et al.* Origin of luminescence from porous silicon deduced by synchrotron light induced optical luminescence. *Nature* **363**, 332–334 (1993).
- Pettifer, R. F. *et al.* X-ray excited optical luminescence (XEOL) study of porous silicon. *Physica B* **208** & **209**, 484–486 (1995).
- Himpel, F. J., Heimann, P., Chiang, T. C. & Eastman, D. E. Geometry-dependent Si(2p) surface core-level excitations for Si(111) and Si(100) surfaces. *Phys. Rev. Lett.* **45**, 1112–1117 (1980).

- Schulz, M. Coulomb energy of traps in semiconductor space-charge regions. *J. Appl. Phys.* **74**, 2649–2657 (1993).
- Read, A. J. *et al.* First principles calculation of the electronic properties of silicon quantum wires. *Phys. Rev. Lett.* **69**, 1232–1235 (1992).
- Tsu, R. & Babic, D. Doping of a quantum dot. *Appl. Phys. Lett.* **64**, 1806–1808 (1994).
- Lehmann, V., Hofmann, F., Moller, F. & Gruning, U. Resistivity of porous silicon—a surface effect. *Thin Film Solids* **255**, 20–22 (1995).
- Pennelli, G. Transient voltage behaviour of free standing porous silicon layers. *J. Appl. Phys.* **80**, 5116–5120 (1996).
- Stauffer, D. & Aharony, A. *Introduction to Percolation Theory* (Taylor & Francis, London, 1994).
- Chen, W. & Ahmed, H. Fabrication of and physics of 2 nm islands for single electron devices. *J. Vac. Sci. Technol. B* **13**, 2883–2887 (1995).
- Fraunhoff, S. T. *et al.* An extended quantum model for porous silicon formation. *J. Electrochem. Soc.* **142**, 615–620 (1995).
- Kanemitsu, Y., Uto, H., Matsumoto, Y., Futagi, T. & Mimura, H. Optical properties of free standing porous films. *Mater. Res. Soc. Proc.* **298**, 265–268 (1993).
- Aldao, C. M., Waddill, P. J., Benning, P. J., Capasso, C. & Weaver, J. H. Photovoltaic effects in temperature-dependent Fermi-level movement for GaAs (110). *Phys. Rev. B* **41**, 6092–6095 (1990).
- Alonso, M., Cimino, R. & Horn, K. Surface photovoltage effects in photoemission from metal-GaP (110) interfaces: importance for band-bending evaluation. *Phys. Rev. Lett.* **64**, 1947–1950 (1990).

Correspondence and requests for materials should be addressed to B.H. (e-mail: hamilton@umist.ac.uk).

Colloid crystal self-organization and dynamics at the air/water interface

H. Hollis Wickman* & Julius N. Korley†

† National Science Foundation, 4201 Wilson Boulevard, Arlington, Virginia 22230 and Center for BioMolecular Science and Engineering, Naval Research Laboratory, Washington DC 20375-5338, USA

† Clark Atlanta University, Atlanta, Georgia 30314, USA

The properties of two-dimensional arrays of micrometre-sized particles are of interest in relation to a wide range of phenomena, including self-organization and phase behaviour in colloid science and condensed-matter physics^{1–3}, the behaviour of dusty plasmas⁴ and the templating of ordered structures for photonic applications⁵. Most studies have used pre-existing particles such as monodisperse latex spheres, which may be manipulated with electric or magnetic fields. In contrast, we report here an inorganic solution that spontaneously precipitates a self-organized two-dimensional colloid crystal at the air/water interface. A solution of calcium hydroxide exposed to air reacts with dissolved carbon dioxide to precipitate microcrystals of calcium carbonate in the form of calcite. These aggregate at the surface to form a disordered gel mat with fractal characteristics⁶. We find, however, that in aged solutions a second population of charged microcrystals with the 'dogtooth spar' morphology appears on the surface. These crystallites, which can be observed by optical microscopy, become organized into a regular triangular lattice. The competition between electrostatic and capillary forces between particles leads to lattice spacings of the order of 125 to 175 μm , 5 to 7 times the diameter of the particles. These structures are stable for around 24 h, but eventually aggregate with the fractal gel. The mechanism of their self-organization, as yet incompletely understood, might provide some insights into similar phenomena in colloid science^{2,3,7–9}.

We prepared solutions of calcium hydroxide with Millipore filtered, deionized water. A 1 ml sample of a 0.3 g l^{-1} , 4×10^{-4} M solution was added to an equal volume of water in a small Petri dish 25 mm diameter by 10 mm tall. Calcium carbonate microcrystals immediately formed in the subphase and diffused to the surface, where they were entrained. Details of the crystal nucleation and growth process are incompletely understood, but are consistent with fast-growing hydrophilic surfaces leading to net hydrophobic facets. The small surface crystallites ($\sim 1\text{--}2\text{ }\mu\text{m}$ diameter) of relative density 2.7 diffuse on the surface and jump into contact via

attractive van der Waals and capillary forces. Ultimately a fractal gel was formed through cluster-cluster aggregation⁶ (Fig. 1a, b). The mineralization and fractal development took about 30 min and were studied in real time using a Leitz-Wetzlar transmission microscope or a Zeiss Axiomat reflection microscope, with objectives ranging from $\times 4$ to $\times 32$.

The new phenomena were observed with solutions that have been stored in closed Nalgene bottles and aged for at least 2 weeks. The bottles were slightly warmed ($35\text{--}45^\circ\text{C}$) in a water bath before dilution as noted above. These solutions yielded fractal-forming calcite crystals that are notably larger ($\sim 10\text{ }\mu\text{m}$) than those in fresh solutions. Further, in the mid-to-late stages (15–30 min) of fractal calcite gel formation, dogtooth crystals began to appear on the water surface. The dogtooth crystals resemble distorted tetrahedra and invariably orient with an apex pointing up, and so appear as a triangle from above.

The structural and chemical identification of the dogtooth crystals is as follows. Figure 2 shows a transmission light microscope view of a segment of a calcite fractal gel mat, decorated with dogtooth crystals (see discussion below), that has been deposited at the bottom of the reaction vessel by evaporation of the mother liquor. Polarized microscopy shows that the dogtooth crystals are nearly optically isotropic along the viewing axis and do not extinguish like rhombohedral calcite. The reagent-grade calcium hydroxide is $>99\%$ calcium (hydroxide or carbonate), according to the manufacturer's (Aldrich Chemical Co.) analysis. However, the level of magnesium and related salts is 0.3 wt%. Because the mass of the dogtooth layers is comparable to this impurity level, the dogtooth and calcite crystals were analysed using electron beam microchemistry. The analyses were indistinguishable in the two cases (D. E. Newbury, personal communication). Electron beam back-scatter pattern analysis of samples such as that shown in Fig. 2 has determined that the crystal structure is calcite, but with a habit consistent with the hexagonal scalenohedron, or 'dogtooth spar' (J. Michael, personal communication). A full report of these analyses is in preparation.

Because plastic bottles can leach monomer, catalyst and so on into solutions under storage conditions, they potentially play a role in the precipitation of the dogtooth crystals. A control experiment was carried out using a Nalgene bottle that had been specially treated to exclude adventitious impurities¹⁰ (we thank B. Rappoli for assistance). The results were the same as with untreated Nalgene. These analyses do not shed light on the nucleation phenomena that

are evidently occurring as the calcium hydroxide solutions age. These solutions are currently being studied using light scattering and other methods to obtain information about the intermediate stages.

Although there are modest variations in the crystal growth, the most dramatic results, described here, were observed with preparations predominantly yielding the dogtooth crystal habit shown in Fig. 2. When first formed, the dogtooth crystals remained isolated and did not aggregate. Pair-wise repulsive collisions were occasionally observed, with a closest approach of roughly five particle diameters. The crystals continued to grow (for an hour or more) until the longest side of the triangle cross-section was $\sim 25\text{ }\mu\text{m}$. The population of crystals in the subphase diminished during this period and eventually disappeared, signalling the end of the visible precipitation reaction. A notable population of hundreds of non-aggregated, dogtooth crystallites was then on the surface, in addition to the fractal gel of rhombohedral calcite crystals, and a few non-aggregated calcite crystals. The gel occupied the centre of the solution surface as a disk $\sim 6\text{ mm}$ in diameter, centred by gravitational forces and surface curvature. Over a period of about 30 min the dogtooth crystals in the annular region between the gel mat and the dish edge organized into a new structure. The crystals first assembled into small patches, with inter-particle separation of $5\text{--}7\text{ }\mu\text{m}$ diameters ($125\text{--}175\text{ }\mu\text{m}$). The patches then coalesced into a large, localized, disk-shaped layer immediately adjacent to the fractal gel (Fig. 1). We call this structure a w-layer by analogy with the self-organized two-dimensional para-crystalline protein structures known as s-layers¹¹ that are found on certain bacterial surfaces. Few free dogtooth crystallites remained; these were distant from the w-layer.

The w-layer acts as a single fluid crystalline unit, and can occupy an area comparable to that of the fractal gel mat. However, the relative masses of the two structures are vastly different. With equal density for the calcite and dogtooth crystals, the mass of the w-layer, derived from microscope images, is at most 0.2 wt% of the total precipitate on the water surface. The lattice order can be high, with more than 50 repeats in a given direction, and is limited by the amount of dogtooth material produced. Not all lattice sites in the w-layer are occupied by a single dogtooth crystal. There are some dyads, triads and so on that are dogtooth crystallites or dogtooth and calcite. At least one dogtooth crystal is normally present at each regular lattice site, sometimes with slightly different growth morphology (truncated apex or apices). Non-dogtooth defect

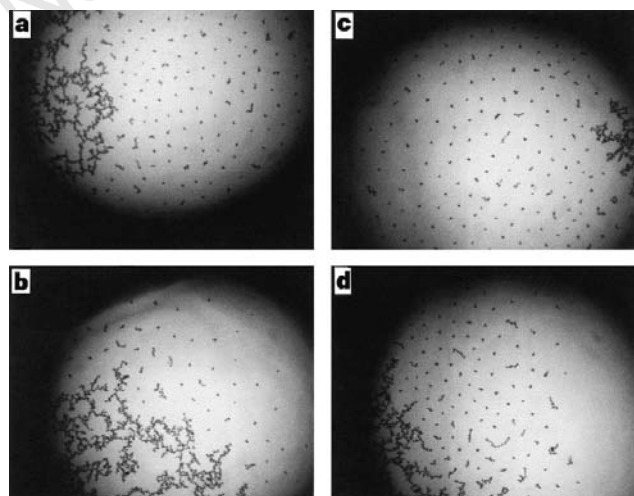


Figure 1 Images of the w-layer. **a**, w-layer at the periphery of the fractal gel at left of image. **b**, Distortion of w-layer to yield a new shape around the periphery of the fractal gel (lower left region); **c**, w-layer re-established; **d**, w-layer beginning to 'absorb' segments from the fractal gel at left in image.

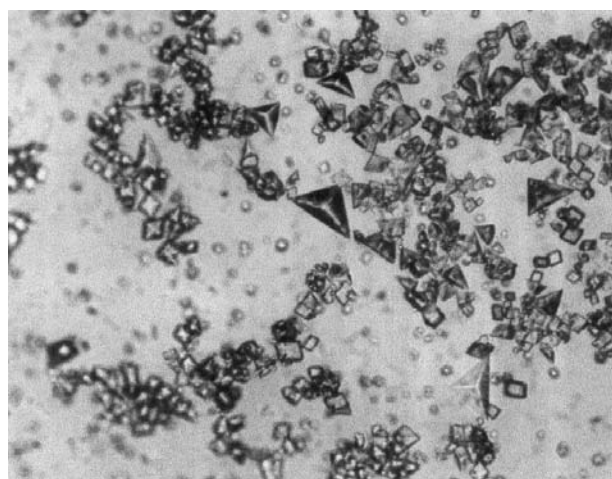


Figure 2 Optical microscope photograph of typical w-layer dogtooth spar (dogtooth) crystals. The long edge of the largest crystal in the centre of the image is $23\text{ }\mu\text{m}$. Rhombohedral calcite crystals are also present. The smallest crystals (not in focus) are calcite on the bottom of the dish.

crystallites distort the lattice and anneal by jumping to an adjacent dogtooth site, at which point the lattice becomes more symmetrical (hexagonal). Owing to gravity, the lattice repeat-distance increases (and rigidity decreases) with radial distance from the edge of the gel mat, but is roughly constant for a given radius. The w-layer responds to disturbances of the air/water interface by hydrodynamic distortions as indicated in Fig. 1. The w-layer can be distorted into an annulus around the periphery of the gel mat (Fig. 1b). If the solution was now left undisturbed for 20 min or so, the w-layer recovered a disk shape (Fig. 1c). One experiment produced nearly monodisperse dogtooth crystals of 25 μm diameter. The w-layer from this solution at one point froze into a rigid hexagonal lattice that was stable for several minutes. Ultimately, an air perturbation caused the lattice to 'melt' into the fluid phase.

Over about 24 h, the w-layer and the fractal gel coalesced in a complex fashion. The w-layer exhibited a surface tension, as indicated in Fig. 1, and this surface tension represents an initial barrier to the immediate aggregation of the charged dogtooth crystallites and the uncharged fractal gel mat. With time, dogtooth crystals occasionally jump and adhere to peripheral pieces of the gel mat. When a small piece of the gel is decorated with 3 or 4 of the dogtooth crystals, the fragment breaks off from the gel. This hybrid cluster than assumes a lattice site in the w-layer (Fig. 1d). In this way, the w-layer converts to a lattice of hybrid clusters. Finally the clusters re-aggregated to form the final fractal gel, which is predominantly calcite but with the periphery decorated with the dogtooth crystallites. After this, no further changes were observed.

The interactions that stabilize the w-layer are thought to result from repulsive electrostatic forces and attractive capillary forces. The charging mechanism of the w-layer particles is under investigation. The electrostatic interactions are mediated primarily through the air. Preliminary experiments involving addition of electrolyte to the subphase did not perturb the w-layer. Lack of rotational correlation between adjacent dogtooth particles argues for monopole interactions. The presence of some dipole or quadrupole interactions cannot be ruled out, because the spatial distribution of charge on the dogtooth particles is unknown. An estimate of the magnitude of the charge is obtained by equating kinetic energy to electrostatic energy during a repulsive collision. The result is about 50 electron charges on a typical tetrahedral crystallite. The presence of capillary forces is signalled by water surface curvature at the periphery of the tetrahedral particles, which can be seen by microscopy. The functional forms of the attractive inverse-distance term^{12,13} and repulsive scaled exponential term¹⁴ are similar in form but opposite in sign to the usual DLVO (Derjaguin–Landau–Verwey–Overbeek) situation¹⁵. The sum of the interactions in the DLVO case can yield an interaction potential with a secondary minimum. By analogy (but with a greatly scaled interparticle separation) a similar minimum is postulated as a primary factor in the stabilization of the w-layer.

Implicit in our discussion is the notion that the air/water interface is required for the stability of the w-layer. Evidence for this is that small calcite crystals are always found on the bottom of the reaction vessel, but the large dogtooth crystals are observed only at the air/water interface. If there is an intrinsic role for the air/water interface beyond the mechanical entrainment of charged particles, the w-layer situation may bear some relations to self-organization of micrometre-size charged spheres near surfaces as reported by Larson and Grier⁷ and Kepler and Fraden⁸. In such cases, there is evidence for attractive interactions beyond the classical DLVO model, although the origin is incompletely understood⁹. □

Received 22 September 1997; accepted 4 March 1998.

1. Skeltorp, A. T. One- and two-dimensional crystallization of magnetic holes. *Phys. Rev. Lett.* **51**, 2306–2309 (1983).
2. Murray, C. A. & Van Winkle, D. H. Experimental observation of two-stage melting in a classical two-dimensional screened Coulomb system. *Phys. Rev. Lett.* **58**, 1200–1203 (1986).
3. Murray, C. A. & Grier, D. G. Video microscopy of monodisperse colloidal systems. *Annu. Rev. Phys. Chem.* **47**, 421–462 (1996).

4. Thomas, H. *et al.* Plasma crystal: Coulomb crystallization in a dusty plasma. *Phys. Rev. Lett.* **73**, 652–655 (1994).
5. Van Blaaderen, A. & Wiltzius, P. Template-directed colloidal crystallization. *Nature* **385**, 321–324 (1997).
6. Nakayama, T., Nakahara, A. & Matsushita, M. Cluster-cluster aggregation of calcium carbonate colloid particles at the air/water interface. *J. Phys. Soc. Jpn* **64**, 1114–1119 (1996).
7. Larson, A. E. & Grier, D. G. Like-charge attractions in metastable colloidal crystallites. *Nature* **385**, 230–233 (1997).
8. Kepler, G. M. & Fraden, S. Attractive potential between confined colloids at low ionic strength. *Phys. Rev. Lett.* **73**, 356–359 (1994).
9. Murray, C. A. When like charges attract. *Nature* **385**, 203–204 (1997).
10. Moody, J. R. & Lindstrom, R. M. Selection and cleaning of plastic containers for storage of trace element samples. *Anal. Chem.* **77**, 2264–2267 (1977).
11. Sleytr, U. B. & Messner, P. Crystalline bacterial cell surface layers (s-layers). *Ency. Microbiol.* **1**, 605–613 (1992).
12. Medina-Noyola, M. & Ivlev, B. I. Interaction in colloidal systems: bucking and melting. *Phys. Rev. E* **52**, 6281–6288 (1995).
13. Camion, C., Roussel, J. F., Faure, R. & Blanc, R. Mesure des forces d'attraction entre spheres partiellement immergées: Influence des interfaces. *Europhys. Lett.* **3**, 449–457 (1987).
14. Chan, D. Y. C., Henry, J. D. Jr & Whie, J. R. The interaction of colloidal particles collected at fluid interfaces. *J. Colloid Interface Sci.* **79**, 410–418 (1981).
15. Israelachvili, J. N. *Intermolecular and Surface Forces* (2nd edn) (Academic, London, 1992).

Acknowledgements. The hospitality of the Naval Research Laboratory and partial support from the Office of Naval Research during a sabbatical visit of H.H.W. are gratefully acknowledged. We thank K. Hathaway, B. Gaber, P. Schoen, D. Turner and G. Lee for useful discussions and access to instruments used in the study. J.N.K. was a summer student whose visit to the Naval Research Laboratory was supported by an Undergraduate Minority Student Program Grant from the Chemical and Transport Systems Division, National Science Foundation.

Correspondence and requests for materials should be addressed to H.H.W. (e-mail: hwickman@nsf.gov).

Continuing decline in the growth rate of the atmospheric methane burden

E. J. Dlugokencky*, K. A. Masarie*†, P. M. Lang* & P. P. Tans*

* NOAA Climate Monitoring and Diagnostics Laboratory, 325 Broadway, Boulder, Colorado 80303, USA

† Cooperative Institute for Research in Environmental Sciences, University of Colorado, Boulder, Colorado 80309, USA

The global atmospheric methane burden has more than doubled since pre-industrial times^{1,2}, and this increase is responsible for about 20% of the estimated change in direct radiative forcing due to anthropogenic greenhouse-gas emissions. Research into future climate change and the development of remedial environmental policies therefore require a reliable assessment of the long-term growth rate in the atmospheric methane load. Measurements have revealed that although the global atmospheric methane burden continues to increase² with significant interannual variability^{3,4}, the overall rate of increase has slowed^{2,5}. Here we present an analysis of methane measurements from a global air sampling network that suggests that, assuming constant OH concentration, global annual methane emissions have remained nearly constant during the period 1984–96, and that the decreasing growth rate in atmospheric methane reflects the approach to a steady state on a timescale comparable to methane's atmospheric lifetime. If the global methane sources and OH concentration continue to remain constant, we expect average methane mixing ratios to increase slowly from today's 1,730 nmol mol⁻¹ to ~1,800 nmol mol⁻¹, with little change in the contribution of methane to the greenhouse effect.

This study is based on CH₄ measurements from discrete, whole-air samples collected at a subset of the NOAA Climate Monitoring and Diagnostics Laboratory (CMDL) cooperative air sampling network sites. Samples are obtained in pairs approximately weekly from 29 land-based sites ranging in latitude from 82° N to 90° S, and 1 sample pair about every 1.5 weeks at 5° latitude intervals between 30° N and 35° S from 2 ships in the Pacific Ocean. The data set is dominated by samples that are representative of the remote marine boundary layer. All samples were analysed for CH₄ at CMDL by gas

chromatography with flame ionization detection. Methane dry-air mole fractions (nmol mol^{-1}) are assigned by referencing each aliquot of sample to an internally consistent methane standard scale. The average relative precision of the measurements is $\sim 0.2\%$. The excellent precision and internally consistent standard gas scale ensure that small changes in the methane trend can be identified. Details of the sampling and analysis procedures are available elsewhere,²⁵ (the data are available at [ftp.cmdl.noaa.gov path: ccg/ch4](http://ftp.cmdl.noaa.gov/path/ccg/ch4)).

We focus our analysis on globally and zonally averaged time series that are based on spatial and temporal smoothing of the measurements. First, the average trend and seasonal cycle for each time series are approximated by fitting a second-order polynomial and four harmonics to the data:

$$f(t) = a_1 + a_2 t + a_3 t^2 + \sum_{i=1}^4 [a_{2i+2} \sin(2\pi i t) + a_{2i+3} \cos(2\pi i t)] \quad (1)$$

Deviations from the average trend and seasonal cycle described by equation (1) are accounted for by first transforming the residuals from the fit to the frequency domain using a fast Fourier-transform algorithm, and then multiplying them by a low-pass filter⁶. The filter function is designed to have a transmission of 0.5 at a specified cut-off frequency and nearly zero transmission at twice the cut-off frequency. Two cut-off values are used. A cut-off frequency of $0.55 \text{ cycles yr}^{-1}$ is used to determine interannual variations in the trend. This result is added to the polynomial component of equation (1) to give a deseasonalized trend function, $F(t)$. A second cut-off frequency of $4.56 \text{ cycles yr}^{-1}$ is used to capture

interannual variations in the average seasonal cycle. This result is added to equation (1) to give a smooth curve fitted to the data, $G(t)$. Global and zonal averages are calculated by extracting synchronized values from each $G(t)$, and, at each time step (48 steps per year), another curve is fitted as a function of latitude using a digital filtering technique⁷. This time-series of CH_4 latitude gradients defines a surface of zonally averaged CH_4 mole fraction in the boundary layer from which weighted averages are calculated⁵ at ~ 8 -day intervals for global, hemispheric and semihemispheric zones. Uncertainties in zonally averaged parameters are estimated with a nonparametric statistical technique²⁻⁵.

Globally averaged CH_4 mole fractions are plotted in Fig. 1a (filled circles). The solid line is the deseasonalized trend, $F(t)$, and its derivative is plotted in Fig. 1b. The average global trend from 1984 to 1996 was $(8.9 \pm 0.1) \text{ nmol mol}^{-1} \text{ yr}^{-1}$, but the rate of growth decreased by $(0.9 \pm 0.1) \text{ nmol mol}^{-1} \text{ yr}^{-2}$. Global and zonally averaged growth rates are summarized in Table 1. Steele *et al.*⁵ found that the CH_4 growth rate in the high Northern Hemisphere (30 – 90°N) during 1984–90 was decreasing ~ 3 times faster than in the other semihemispheres. With the addition of measurements from 1991 to 1996, there are now only small differences among trends and changes in trends calculated for each semihemisphere. This is expected because the CH_4 atmospheric lifetime is ~ 9 years, and CH_4 is relatively well mixed in the atmosphere, which does not allow mixing ratios in different places to continue diverging for many years.

The instantaneous growth rate curve plotted in Fig. 1b indicates that the trend decreased from $\sim 14 \text{ nmol mol}^{-1} \text{ yr}^{-1}$ in 1984 to

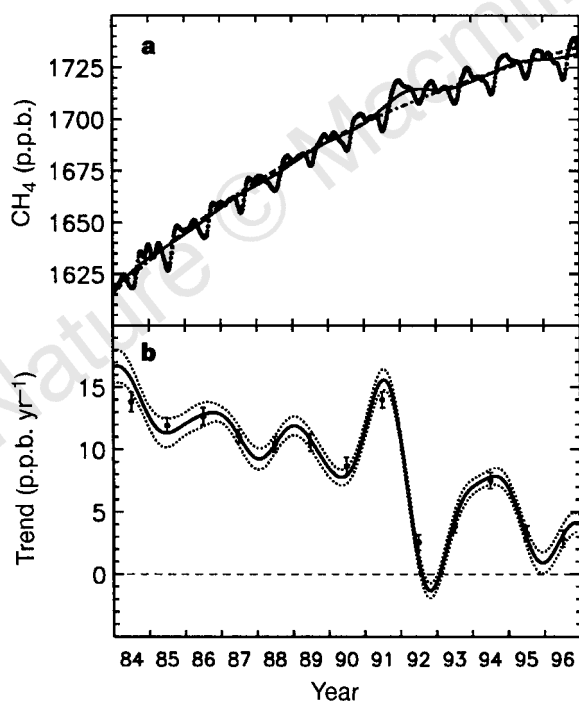


Figure 1 Result of spatial and temporal smoothing of CH_4 measurements from CMDL air sampling network sites. **a**, Globally averaged CH_4 mole fractions (where p.p.b. is used to abbreviate nmol mol^{-1}), plotted as solid symbols. The solid line is a deseasonalized trend curve fitted to the data. The dashed-dot line is a fit of equation (4) to the trend line calculated using the coefficients $[\text{CH}_4]_{\text{ss}} = 1,779 \text{ nmol mol}^{-1}$, and $\tau = 10.0 \text{ yr}$. **b**, Atmospheric CH_4 instantaneous growth rate is plotted as the solid line. This curve is the derivative with respect to time of the trend curve in **a**. The uncertainties (dotted lines) are determined with a non-parametric statistical technique. The symbols are annual increases in CH_4 determined from the deseasonalized trend line ($F(t)$ in **a**); their uncertainties were also determined with the non-parametric statistical method.

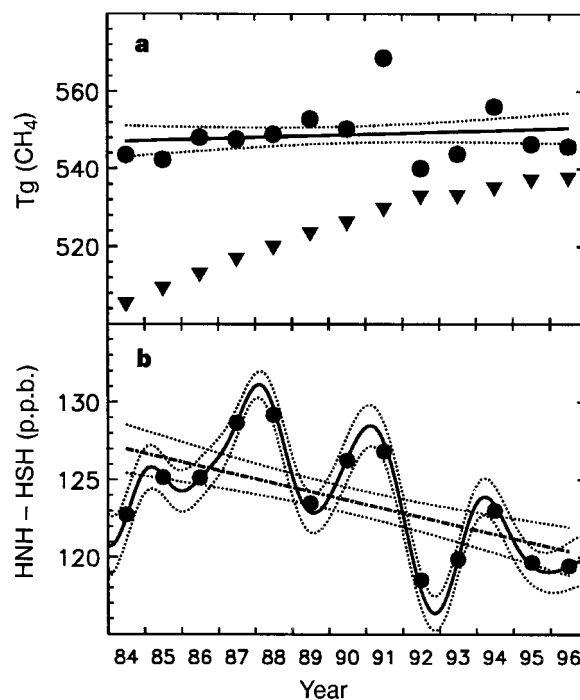


Figure 2 Results of calculations of CH_4 source and sink from mass-balance equation and intersemihemispheric difference are plotted as a function of time. **a**, Global CH_4 source (circles) and sink (triangles) calculated with equation (3) from CMDL measurements of the global burden and annual increase. A straight line fitted to the annual source terms gives a slope of $0.3 \pm 0.6 \text{ Tg yr}^{-1}$. **b**, Annual (circles) and weekly (curved solid line) differences between zonally averaged CH_4 mole fractions in the latitude zones 30 – 90°N (HNH) and 30 – 90°S (HSH). The average difference is $123.6 \text{ nmol mol}^{-1}$, and the trend in the difference (dotted-dashed line) is $-0.6 \pm 0.2 \text{ nmol mol}^{-1} \text{ yr}^{-1}$. Uncertainties in the weekly differences are determined with the bootstrap technique.

Table 1 Average CH₄ growth rate and change in growth rate for 1984–96

Zone	Trend (nmol mol ⁻¹ yr ⁻¹)	ΔTrend (nmol mol ⁻¹ yr ⁻²)
Global	8.9 ± 0.1	–0.9 ± 0.1
N. Hemisphere	8.8 ± 0.1	–1.0 ± 0.1
S. Hemisphere	9.0 ± 0.1	–0.8 ± 0.1
30–90° N	8.6 ± 0.2	–1.1 ± 0.1
Eq.–30° N	9.1 ± 0.2	–1.0 ± 0.1
Eq.–30° S	8.9 ± 0.1	–0.8 ± 0.1
30–90° S	9.1 ± 0.1	–0.8 ± 0.1

Data for global and zonal averages are shown. Uncertainties are estimated using a bootstrap method: 100 bootstrap air sampling networks are produced by randomly picking sites from the CMDL network, with restitution, so that each site may be repeated while some sites are omitted. The data from each bootstrap network were analysed to produce 100 zonally smoothed data sets. Each of these was fitted by equation (1). For the average trend in this table, slopes from linear fits were used. ΔTrend was determined from twice the coefficient a_3 in equation (1). The uncertainties stated are 1σ of the 100 coefficients determined from 100 fits to the zonally averaged data sets.

~3 nmol mol⁻¹ yr⁻¹ in 1996. This decrease has not been monotonic; significant interannual variability in the growth rate is apparent in Fig. 1, particularly after the eruption of Mt Pinatubo in 1991. Significant increases in CH₄ and CO growth rates were observed in the tropics and high southern latitudes soon after the eruption. Calculations from a radiative transfer model³ showed that ultraviolet actinic flux in the wavelength region 290–310 nm was attenuated by ~12% immediately after the eruption owing to absorption by SO₂, and that the ultraviolet flux was perturbed for up to one year because of scattering by sulphate aerosols. The main sink for tropospheric CO and CH₄ is reaction with the OH radical, whose production depends strongly on the photolysis of ozone by solar ultraviolet radiation. Dlugokencky *et al.*³ have therefore attributed the increased CH₄ and CO growth rates to decreased [OH], resulting in smaller sinks. A period of net decrease in globally averaged methane levels was observed in late 1992 and early 1993. Zonal averages calculated from our measurements suggest this was driven in large part by low CH₄ growth rate in the high latitudes of the Northern Hemisphere, possibly related to decreased emissions from fossil fuel in the former Soviet Union⁴ and to decreased emissions from boreal wetlands resulting from cooler temperatures after the eruption of Mt Pinatubo⁸. An increase in the tropical CH₄ sink due to decreased stratospheric O₃ (ref. 9) and decreased tropical emissions from biomass burning¹⁰ may also help explain the decrease in global growth rate.

It remains important to try to understand the reason for the long-term decrease in growth rate, because the trend is an important constraint on the global CH₄ budget and it affects climate change policies. The change in the global burden of CH₄ is given by:

$$d[\text{CH}_4]/dt = Q - [\text{CH}_4]/\tau \quad (2)$$

where [CH₄] is the global CH₄ burden (determined from the CMDL global averages), Q is the sum of all emissions, and τ is the total atmospheric CH₄ lifetime. The term [CH₄]/τ is the sink. Equation (2) can be rearranged to calculate the annual CH₄ source strength:

$$Q = d[\text{CH}_4]/dt + [\text{CH}_4]/\tau \quad (3)$$

The annual increase, d[CH₄]/dt was calculated from $F(t)$. The magnitude and trend of Q are sensitive to τ. We use the lifetime from Prinn *et al.*¹¹, τ = 8.9 yr, which is based on analysis of atmospheric methylchloroform measurements. They found no significant trend in [OH] (0 ± 0.2% yr⁻¹) for the period 1978–94, and, therefore, no trend in CH₄ lifetime. Dlugokencky *et al.*¹² found that trends in the peak-to-peak seasonal cycle amplitude of atmospheric methane were consistent with the trend in amplitude expected due to the increasing atmospheric burden of CH₄, but not with a change in globally averaged [OH]. Model studies suggest that feedbacks on OH caused by the increasing atmospheric burden of CH₄ should lead to an increased CH₄ lifetime¹³. The indirect determination of [OH] by Prinn *et al.*¹¹ suggests the existence of

factors compensating for the increase in CH₄ burden, at least for the relatively small time period of high-precision atmospheric measurements. Soil processes make only a small contribution to the sink, which is addressed later. As the CH₄ lifetime is assumed constant during our measurements, interannual variability in OH is accounted for in the source term. Therefore, Q is a pseudo-source term, and it includes deviations from the mean atmospheric lifetime. We converted our measurements from mole fraction to mass in Tg (where 1 Tg = 10¹² g) with the conversion factor 1 nmol mol⁻¹ = 2.767 Tg (ref. 14).

The annual global source, Q, and the sink calculated from our global averages are plotted in Fig. 2a. Through the period of our measurements, the annual mean of Q (±1σ) was 549 ± 7 Tg, in excellent agreement with other CH₄ budget studies^{14,15}. There are significant interannual variations in Q, but a straight line fitted to Q gives a slope of 0.3 ± 0.6 Tg yr⁻¹, suggesting no significant trend in our pseudo-source from 1984 to 1996. The increase in the CH₄ sink shown in Fig. 2a is due to the increasing CH₄ burden (last term in equation (3)). Therefore, the long-term decrease in CH₄ growth rate can be explained as an approach to steady-state where total source strength and [OH] have been nearly constant. We repeated this analysis using lifetimes of 11 and 8.4 years. The shorter lifetime was used to simulate a small (~30 Tg yr⁻¹) soil sink (IPCC recommends τ = 8.6 ± 1.6 yr)²⁰. The annual means and trends in CH₄ source strength obtained were 449 ± 7 Tg and 0.2 ± 0.6 Tg yr⁻¹ for τ = 11 yr, and 580 ± 8 Tg and 0.5 ± 0.6 Tg yr⁻¹ for τ = 8.4 yr. This shows that our analysis is sensitive to the methane atmospheric lifetime, but even with a relatively short τ = 8.4 yr, the relative rate of increase in source strength is less than 0.1% yr⁻¹.

Using the above result—that CH₄ emissions and [OH] are nearly constant—we can derive a kinetic expression that yields the steady-state CH₄ mole fraction and shows why CH₄ continues to increase while its emissions and lifetime remain constant. At steady-state, d[CH₄]/dt = 0, and Q = [CH₄]_{ss}/τ, where the subscript 'ss' indicates steady state. The approach of CH₄ to steady-state for this system, with a zeroth-order source and pseudo-first-order loss of CH₄ (obtained by making these substitutions into equation (2) and solving), is given by:

$$[\text{CH}_4](t) = [\text{CH}_4]_{ss} - ([\text{CH}_4]_{ss} - [\text{CH}_4]_0)\exp[-t/\tau] \quad (4)$$

[CH₄]₀ is the global CH₄ burden on 1 January 1984. The time constant for approach to steady-state is the CH₄ lifetime. Here, [CH₄]_{ss}, ([CH₄]_{ss} - [CH₄]₀), and τ are determined from a nonlinear curve fitting routine. Equation (4) was fitted to the deseasonalized global averages in Fig. 1a; a steady-state CH₄ mole fraction of 1,779 nmol mol⁻¹ and a methane lifetime of 10.0 yr are obtained. A curve calculated using these coefficients is shown as a dashed line in Fig. 1a. The lifetime determined by the fitting process (10.0 yr) is slightly longer than the best estimate of the lifetime¹¹ (8.9 yr) because CH₄ emissions probably increased slightly during the measurement period.

The analysis suggests that, unless the CH₄ lifetime is much shorter than ~9 yr, emissions of CH₄ have remained approximately constant from 1984 to 1996. The distribution of emissions may be changing as emissions from some sources decrease while others increase. Such a change in source distribution may be observable as a change in the CH₄ latitude gradient or interhemispheric difference. No significant trend in the difference between Northern and Southern Hemisphere annual averages (–0.2 ± 0.2 nmol mol⁻¹ yr⁻¹) was observed, with a mean difference of 89.5 nmol mol⁻¹. A larger trend of –0.6 ± 0.2 nmol mol⁻¹ yr⁻¹, with a mean difference of 123.6 nmol mol⁻¹, was observed in the difference between zonal averages calculated for 30–90° N and 30–90° S (Fig. 2b). This trend suggests that small changes in the distribution of sources may have occurred while the total source strength has remained approximately constant. Changes in the distribution of CH₄ emissions may be accompanied by changes in the distribution of NO_x and CO

emissions. Gupta and Cicerone¹⁶ found that in a two-dimensional model simulation where 10% of CH₄, CO and NO_x emissions were shifted from northern mid-latitudes to the northern tropics, the CH₄ steady-state lifetime decreased by 1.4%.

The main sources of CH₄ to the atmosphere have been identified¹⁴, but emission rates from individual sources are not known well enough to quantify trends in sources. Despite this, evidence exists that some regional sources may be changing. In the states of the former Soviet Union, production of natural gas, oil and coal have all declined during the late 1980s and early 1990s (<http://www.bp.com/bpstats>), and the fraction of gas that is vented or flared has also decreased¹⁷. This would lead to decreased emissions from the high northern latitudes. If this speculative scenario is true, emissions from the tropics must have increased to maintain a nearly constant global source strength.

Our conclusions that CH₄ sources have been nearly constant from 1984 to 1996 has important implications for policy. If CH₄ sources and [OH] remain at 1996 levels, the globally averaged CH₄ mole fraction will slowly increase to ~1,800 nmol mol⁻¹ over the next two decades, and then stabilize. As the imbalance between CH₄ sources and sinks is currently very small, policies that reduce CH₄ emissions by a few per cent would nearly ensure a decrease of CH₄ levels in the atmosphere¹⁸. We point out, however, that quantifying small, long-term trends in specific CH₄ source strengths remains impossible, so the future atmospheric burden of CH₄ cannot be predicted with certainty¹⁹. A better understanding of the CH₄ budget, and how it is changing with time, is still needed to predict more accurately the future role of CH₄ in climate change. □

Received 9 September 1997; accepted 24 March 1998.

- Etheridge, D. M., Pearman, G. I. & Fraser, P. J. Changes in tropospheric methane between 1841 and 1978 from a high accumulation rate Antarctic ice core. *Tellus* **44**, 282–294 (1992).
- Dlugokencky, E. J., Steele, L. P., Lang, P. M. & Masarie, K. A. The growth rate and distribution of atmospheric methane. *J. Geophys. Res.* **99**, 17021–17043 (1994).
- Dlugokencky, E. J. *et al.* Changes in CH₄ and CO growth rates after the eruption of Mt. Pinatubo and their link with changes in tropical tropospheric UV flux. *Geophys. Res. Lett.* **23**, 2761–2764 (1996).
- Dlugokencky, E. J. *et al.* A dramatic decrease in the growth rate of atmospheric methane in the northern hemisphere during 1992. *Geophys. Res. Lett.* **21**, 45–48 (1994).
- Steele, L. P. *et al.* Slowing down of the global accumulation of atmospheric methane during the 1980's. *Nature* **358**, 313–316 (1992).
- Thoning, K. W., Tans, P. P. & Komhyr, W. D. Atmospheric carbon dioxide at Mauna Loa observatory, 2. Analysis of the NOAA GMCC data, 1974–1985. *J. Geophys. Res.* **94**, 8549–8565 (1989).
- Tans, P. P., Conway, T. J. & Nakazawa, T. Latitudinal distribution of the sources and sinks of atmospheric carbon dioxide derived from surface observations and an atmospheric transport model. *J. Geophys. Res.* **94**, 5151–5172 (1989).
- Hogan, K. B. & Harriss, R. C. Comment on 'A dramatic decrease in the growth rate of atmospheric methane in the northern hemisphere during 1992' by E. J. Dlugokencky *et al.* *Geophys. Res. Lett.* **21**, 2445–2446 (1994).
- Bekki, S., Law, K. S. & Pyle, J. A. Effect of ozone depletion on atmospheric CH₄ and CO. *Nature* **371**, 595–597 (1994).
- Lowe, D. C., Brenninkmeijer, C. A. M., Brailsford, G. W., Lassey, K. R. & Gomez, A. J. Concentration and ¹³C records of atmospheric methane in New Zealand and Antarctica: Evidence for changes in methane sources. *J. Geophys. Res.* **99**, 16913–16925 (1994).
- Prinn, R. G. *et al.* Atmospheric trends and lifetime of CH₃CCl₃ and global OH concentrations. *Science* **269**, 187–192 (1995).
- Dlugokencky, E. J., Masarie, K. A., Tans, P. P., Conway, T. J. & Xiong, X. Is the amplitude of the CH₄ seasonal cycle changing? *Atmos. Environ.* **31**, 21–26 (1997).
- Prather, M. J. Time scales in atmospheric chemistry: Theory, GWPs for CH₄ and CO, and runaway growth. *Geophys. Res. Lett.* **23**, 2597–2600 (1996).
- Fung, I. *et al.* Three-dimensional model synthesis of the global methane cycle. *J. Geophys. Res.* **96**, 13033–13065 (1991).
- Cicerone, R. J. & Oremland, R. S. Biogeochemical aspects of atmospheric methane. *Global Biogeochem. Cycles* **2**, 299–327 (1988).
- Gupta, M. & Cicerone, R. Perturbation to global tropospheric photochemistry due to changes in latitudinal distributions of surface sources of CH₄, CO, and NO_x. *Eos* **78**, 90 (1997).
- Law, K. S. & Nisbet, E. G. Sensitivity of the CH₄ growth rate to changes in CH₄ emissions from natural gas and coal. *J. Geophys. Res.* **101**, 14387–14397 (1996).
- Hogan, K. B., Hoffman, J. S. & Thompson, A. M. Methane on the greenhouse agenda. *Nature* **354**, 181–182 (1991).
- Khalil, M. A. K. & Rasmussen, R. Decreasing trend of methane: unpredictability of future concentrations. *Chemosphere* **26**, 803–814 (1993).
- Schimmel, D. *et al.* in *Intergovernmental Panel on Climate Change (IPCC), Climate Change 1995, The Science of Climate Change* 94 (Cambridge University Press, Cambridge, 1996).
- Interactive Data Language Research Systems (Boulder, Colorado, 1997).

Acknowledgements. We thank all agencies that have assisted us with the cooperative air sampling network, and Blue Star Line for their continued support of our sampling efforts. We are grateful for the efforts of all network observers, and thank T. Conway and R. Cicerone for comments. This work was supported in part by the Atmospheric Chemistry Project of the NOAA Climate and Global Change Program and the US Environmental Protection Agency.

Correspondence and requests for materials should be addressed to E.J.D. (e-mail: edlugokencky@cmdl.noaa.gov).

Influence of volcanic eruptions on Northern Hemisphere summer temperature over the past 600 years

K. R. Briffa*, P. D. Jones*, F. H. Schweingruber† & T. J. Osborn*

* Climatic Research Unit, University of East Anglia, Norwich NR4 7TJ, UK

† Swiss Federal Institute of Forest, Snow and Landscape Research, Zürcherstrasse 111, CH-8903 Birmensdorf, Switzerland

A network of temperature-sensitive tree-ring-density chronologies provides circum-hemisphere information on year-by-year changes in summer warmth in different regions of the northern boreal forest¹. Combining these data into a single time-series provides a good summer-temperature proxy for northern high latitudes and the Northern Hemisphere as a whole². Here we use this well dated, high-resolution composite time-series to suggest that large explosive volcanic eruptions produced different extents of Northern Hemisphere cooling during the past 600 years. The large effect of some recent eruptions is apparent, such as in 1816, 1884 and 1912, but the relative effects of other known, and perhaps some previously unknown, pre-nineteenth-century eruptions are also evaluated. The most severe short-term Northern Hemisphere cooling event of the past 600 years occurred in 1601, suggesting that either the effect on climate of the eruption of Huaynaputina, Peru, in 1600 has previously been greatly underestimated, or another, as yet unidentified, eruption occurred at

Table 1 Years of extreme low tree-ring density averaged around the Northern Hemisphere

NHD1*		σ Value†	NH anom.‡	NHD2	
Rank	Year		(°C)	Rank	Value
1	1601	-6.90	-0.81	1	-6.75
2	1816	-4.33	-0.51	7	-3.36
3	1641	-4.31	-0.50	2	-4.27
4	1453	-4.24	-0.50	3	-4.05
5	1817	-3.76	-0.44	10	-3.18
6	1695	-3.50	-0.41	4	-3.64
7	1912	-3.33	-0.39	6	-3.38
8	1675	-3.13	-0.37	92	-1.50
9	1698	-3.08	-0.36	5	-3.40
10	1643	-2.99	-0.35	22	-2.43
11	1699	-2.96	-0.35	20	-2.50
12	1666	-2.89	-0.34	11	-3.16
13	1884	-2.89	-0.34	12	-3.09
14	1978	-2.80	-0.33	14	-2.89
15	1837	-2.78	-0.32	9	-3.23
16	1669	-2.77	-0.32	67	-1.76
17	1587	-2.64	-0.31	19	-2.50
18	1740	-2.61	-0.30	76	-1.72
19	1448	-2.57	-0.30	47	-2.01
20	1992	-2.56	-0.30	52	-1.92
21	1836	-2.48	-0.29	24	-2.40
22	1818	-2.45	-0.29	8	-3.32
23	1495	-2.42	-0.28	36	-2.22
24	1968	-2.38	-0.28	29	-2.32
25	1742	-2.35	-0.27	17	-2.64
26	1783	-2.35	-0.27	18	-2.51
27	1667	-2.35	-0.27	204	-0.82
28	1642	-2.22	-0.26	13	-2.94
29	1819	-2.21	-0.26	25	-2.36
30	1446	-2.20	-0.26	117	-1.31

*The ranked lowest 5% of values in the series made up as the average of five large regional series (see Fig. 2). Bold values also occur in the bottom 5% of a series formed as the simple average of all sites (NHD2).

†σ values are expressed as standardized departures from the mean for AD 1881–1960.

‡Estimated NH temperature anomaly (AD 1881–1960 base). See Methods section for further details.

the same time. Other strong cooling events occurred in 1453, seemingly confirming a 1452 date for the eruption of Kuwae, southwest Pacific, and in 1641/42, 1666, 1695 and 1698.

A number of published regional dendroclimatic temperature reconstructions have made passing reference to the coincidence between certain localized cold summers and the dates of various volcanic eruptions^{3,4}, and some potential for identifying large volcanic signals has been identified in a combined set of North American and European chronologies⁵. Recently, the network of wood-density chronologies has expanded greatly and now encompasses the northern boreal forest regions of Siberia so that, for the first time, it spans the entire Northern Hemisphere, currently incorporating replicated series from over 380 locations.

The tree-ring series that constitute each site chronology were detrended to remove possible bias associated with changes in the age of the sample trees, so that each chronology and the overall mean record is expected to preserve interannual and decadal timescales of temperature variability with greatest fidelity (see Methods section). It is generally accepted that large-scale temperature variability on just these timescales is influenced by the frequency and magnitude of large volcanic eruptions⁶. However, the extent to which the strength of this influence can be tested beyond the confines of the twentieth century is limited by the poor availability of volcanic and climate histories which, of necessity, should be highly resolved and precisely dated^{7,8}. We examine the evidence for volcanic forcing of Northern Hemisphere (NH) summer temperatures by comparing our average NH tree-ring-density record with previous proxies of large-scale volcanic activity and with the historical record of large eruptions over the past 600 years. We do not assume that these tree-ring data represent surface temperature perfectly, or that all low-density values are forced by volcanic eruptions⁵. However, we do consider that the breadth of temperature sensitivity and firm dating control in our data provide as rigorous a basis for examining the nature of a volcano/temperature link over recent centuries as any achieved to date.

Two methods of using the site density series to construct a mean NH record have been explored. In each, time-dependent changes in

site spatial coverage are accounted for by adjusting the variance, in accord with the diminishing number of sites, especially in the early data (see Methods section). The two derived records are, nevertheless, highly correlated ($r = 0.92$, AD 1400–1990) and similar extreme values are indicated in the lowest 5% of the distribution of each of them (Table 1). The (preferred) NH density record, NHD1 (Fig. 1), formed as the average of 5 subcontinental-scale mean density series (Fig. 2), correlates with a series of instrumental summer (April–September) average temperatures⁹ combined over the areas corresponding to the chronology sites² at $r = 0.76$ (AD 1881–1960; $p \ll 0.01$ after correcting degrees of freedom for autocorrelation). The correlation for the same summer season with temperatures averaged over the whole Northern Hemisphere (land and marine regions) is $r = 0.57$. We note that after 1960, there is a reduction in the decadal-timescale correspondence between the instrumental temperature records and the density series, possibly reflecting some large-scale anthropogenic influence causing the densities to decline².

The most representative syntheses of past large-scale volcanic activity are generally recognised as the dust veil index (DVI)¹⁰, the volcanic explosivity index (VEI)¹¹ and the ice-core volcanic index (IVI)¹². Each emphasises a different aspect of eruption activity. All have some inadequacy in spatial and temporal representation and some imprecision in resolution and dating. The IVI, which quantifies widespread ice acidity and the DVI, should contain a significant temperature 'signal'. The VEI, which measures only the magnitude of eruptions and takes no account of likely atmospheric aerosol production, might be expected to show the lowest correlation with temperature¹¹. The NH summer temperature correlations with these series (DVI, -0.37 ; VEI, -0.35 ; IVI, -0.33 ; AD 1881–1960) are all lower than for NHD1.

The correlations between the yearly values of NHD1 (truncated above 0) and each of these series are all statistically significant ($p = 0.01$ or better). (DVI (1500–1990), -0.33 ; VEI and IVI (1400–1990) -0.24 and -0.32 , respectively; see also Supplementary Information). Correlations based on non-overlapping five-year averages (hence reducing the effect of delayed spatial responses

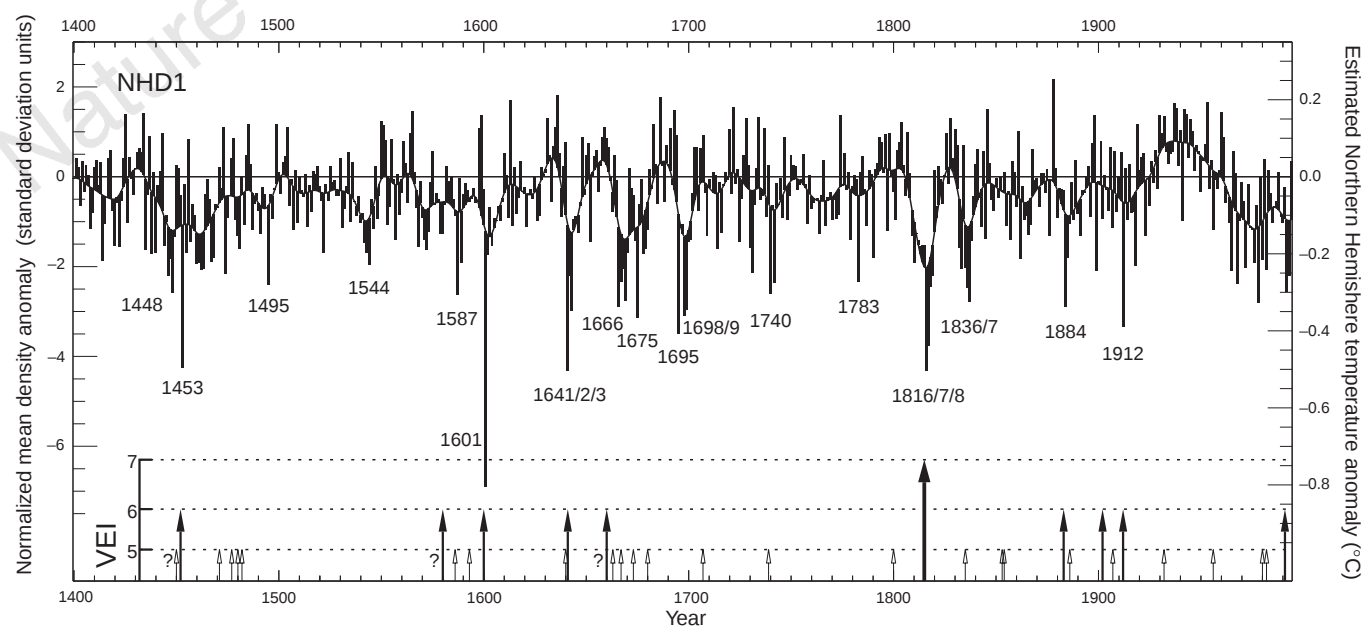


Figure 1 The NHD1 average of NH tree-ring-density chronologies. Values are shown as standardized anomalies from the AD 1881–1960 period (left-hand axis) and in the form of regression-based estimates of NH mean temperature anomalies (right-hand axis). The dates of a number of extreme low values are indicated (see Table 1). The curve shows bi-decadal smoothed values. The main

explosive eruptions (VEI 5, 6 and 7) listed in ref. 14 are indicated on the lower axis. Question marks indicate uncertainty because dates are based on radiocarbon. These data and those for NHD2 (see Methods section) are available from the NOAA NGDC palaeoclimate database (<http://www.ngdc.noaa.gov/paleo.html>).

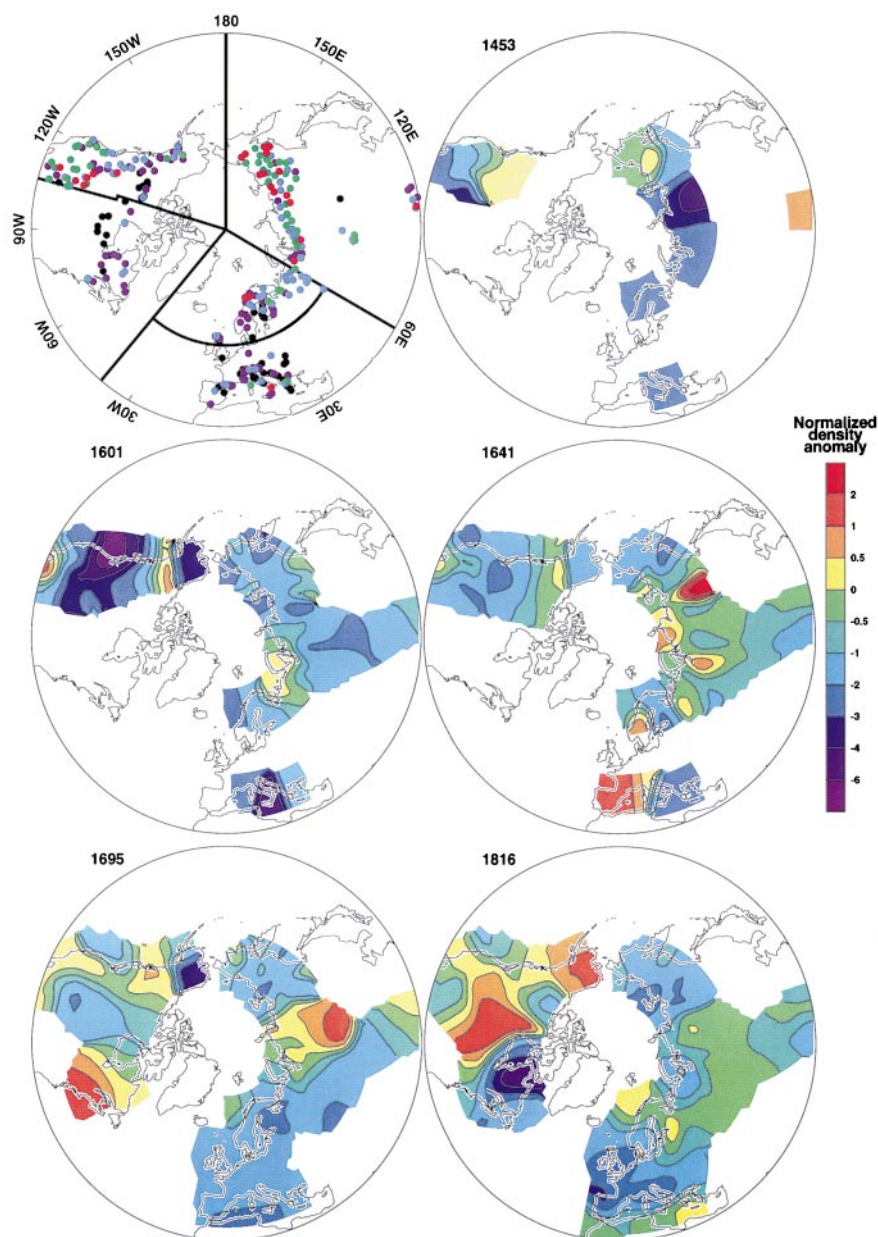


Figure 2 Selected spatial patterns of tree-ring density. Top left, changing spatial coverage of the tree-ring densitometric network over the past five centuries. Red dots denote chronologies starting before AD 1500; green, 1600; blue, 1700; purple, 1800; and black, 1900. NHD1 was constructed as the average of five mean chronologies comprising data from the regions separated by black lines. The other maps show normalized chronology density (with respect to AD 1881–1960) in selected years of low overall mean density (see Fig. 1). Negative values indicate regions of reduced summer temperature.

and slight dating uncertainties, especially in the IVI) are all higher (−0.46, −0.33 and −0.45, respectively) and higher again if the more recent, and hence better defined, five-year averages since 1600 are compared (−0.51, −0.43 and −0.52, respectively).

The largest historically documented eruption of the modern (instrumental) era is that of Tambora, Indonesia, in April 1815¹³. This is one of only four Holocene eruptions to have been assigned a VEI of 7 (ref. 14; see also Table 2). Instrumental, historical and various tree-ring evidence all show widespread cold conditions, especially in eastern North America and western Europe in 1816¹³ and the various ice core data^{7,15–20} invariably show a strong acidity signal associated with this year. Very low summer temperatures (Fig. 2) are indicated in our data in 1816 (rank second lowest; compare Table 1) but also in 1817 (tenth) and 1818 (eighth). The decade of the 1810s is shown to be the coldest in the record, due in part to a prior, as yet still unidentified, eruption in 1809 originally indicated by ice-core evidence²¹ and later shown in regional tree-ring data^{3,4}. This also shows up as a low value in 1810 (15th) in our NHD2 record (it is 49th in NHD1). Evidence of other, more recent, climatically effective volcanoes, shown in NHD1 by low values in

1912 (7th), 1884 (13th), 1836/7 (21st/15th) and even 1992 (20th), can be ascribed to the eruptions of Novarupta, Krakatau, Cosiguina and Pinatubo respectively, all of which, except Cosiguina²², have VEI of 6 (see Table 2). However, the eruption of Santa Maria (1902), also with a VEI of 6, does not show up as very anomalous in our data. We note that this is true also of some of the ice-core records^{7,15,17,18}.

The most extreme value in our record is in 1601 for which we estimate (see Methods) a NH summer anomaly of −0.8 °C (see Figs 1 and 2, and Table 1). This is 0.3 °C cooler than that estimated in 1816. Several previous dendroclimatic reconstructions have indicated the local significance of the 1601 cold anomaly, notably in North America^{4,23}, but it is revealed here as clearly the most important hemispheric tree-ring density event in the past 600 years. Early evidence of a contemporaneous ‘unknown’ eruption came from European historical observations of optical atmospheric phenomena (dim Sun and reddish haze) in 1601 and 1602¹⁰. Subsequent acidity evidence from ice cores in both Greenland^{15,24} and Antarctica^{7,17} indicate that the volcanic aerosol around 1601 was of global extent, and analyses of contemporaneous tephra and dust in Antarctica²⁵ and a tropical ice cap from Quelccaya, Peru²⁶ has

Table 2 Largest explosive volcanic eruptions since AD 1400

Year*	Season†	Volcano and region	Lat. (deg.)	Long. (deg.)	VEI‡
1450C	–	Aniakchak, Alaska	5.7N	158.1W	5(?)
1452	–	Kuwaie, Vanuatu, SW Pacific	16.8S	168.5E	6
1471	3yr	Sakura-Jima, Japan	31.6N	130.7E	5(?)
1477	1	Bardarbunga (Veidivotn), Iceland	64.6N	17.5W	5(?)
1480D	–	St Helens, Washington, US	46.2N	122.2W	5(+)
1482D	–	St Helens, Washington, US	46.2N	122.2W	5
1580C	–	Billy Mitchell, Bougainville, SW Pacific	6.1S	155.2E	6
1586	–	Kelut, Java	7.9S	112.3E	5(?)
1593	–	Raung, Java	8.1S	114.0E	5(?)
1600	1	Huaynaputina, Peru	16.6S	70.9W	6(?)
1640	3	Komaga-Take, Japan	42.1N	140.7E	5
1641	1	Parker, Philippines§	6.1N	124.9E	6
1660C	–	Long Island, New Guinea	5.4S	147.1E	6
1663	3	Usu, Japan	42.5N	140.8E	5
1667	4	Shikotsu (Tarumai), Japan	42.7N	141.3E	5
1673	2	Gamkonora, Halmahera	1.4N	127.5E	5(?)
1680	–	Tongkoko, Sulawesi	1.5N	125.2E	5(?)
1707	1	Fuji, Japan	35.4N	138.7E	5
1739	3	Shikotsu (Tarumai), Japan	42.7N	141.3E	5
1800D	1	St Helens, Washington, US	46.2N	122.2W	5
1815	2	Tambora, Lesser Sunda Is	8.3S	118.0E	7
1835	1	Cosiguina, Nicaragua	13.0N	87.6W	5
1853	1	Chikurachki, Kurile Is	50.2N	155.0E	5(?)
1854	1	Sheveluch, Kamchatka	56.7N	161.4E	5
1883	3	Krakatau, west of Java	6.1S	105.4E	6
1886	3	Okataina (Tarawera), New Zealand	38.1S	176.5E	5
1902	4	Santa Maria, Guatemala	14.8N	91.6W	6(?)
1907	2	Ksudach, Kamchatka	51.8N	157.5E	5
1912	1	Novarupta (Katmai), Alaska	58.3N	155.2W	6
1932	2	Azul, Cerro (Quizapu), Chile	35.7S	70.8W	5(+)
1956	2	Bezymianny, Kamchatka	56.0N	160.6E	5
1980	2	St Helens, US	46.2N	122.2W	5
1982	2	El Chichon, Mexico	17.4N	93.2W	5
1991	3	Pinatubo, Philippines	15.1N	120.4E	6

*C refers to an uncorrected radiocarbon-based estimate; D indicates a dendrochronological date.

† Where known, are indicated as follows: 1, Dec.–Feb. 2, Mar.–May, 3, Jun.–Aug., 4, Sep.–Nov.

‡ Volcanic explosive index as listed in ref. 14. The scale is logarithmic and values of 6 and over are here indicated in bold. In the original ref. 14, (?) indicates a problematic assignment while a (+) indicates an eruption in the upper third of the range for that designation.

§ This eruption is apparently wrongly attributed in ref. 14 to Awu (Indonesia) as the result of an incorrect second-hand report in 1880, and is here given a larger VEI (C. Newhall, personal communication).

identified the source as Huaynaputina, Peru. If the Huaynaputina eruption, dated from documentary sources to 1600²⁶, was the sole cause of the cooling, it would imply a very large sulphur release, though a contemporaneous eruption elsewhere is also possible²⁷. However, the cooling indicated in NHD1 is greater than is suggested by the degree of acidity in the various ice-core records.

The most explosive eruption of the fifteenth century is thought to have been that of Kuwaie, southwest Pacific²⁸. A date of around 1425 (~1420–75) has been proposed on the basis of radiocarbon evidence²⁹. Diverse historical evidence suggests a 1452 date³⁰. Large acid ‘spikes’ in Antarctic^{7,17–19} and (though less pronounced) in Greenland^{15,16} ice cores suggest various dates for the eruption between about 1450 and 1460. Though based on a relatively sparse network (Fig. 2), the 1453 low value in our record (fourth lowest) is exactly coincident with one of the three most notable frost rings identified in high-elevation Californian pines³¹ (the others being in 1601 and 1884) and provides strong support for a 1452 eruption date (allowing a one-year lag for a Southern Hemisphere eruption). Recent data^{24,32} indicate that the sulphur output from Kuwaie may have been the same as for Tambora, giving credence to the extent of cooling indicated by our data.

There is a concentration of apparently cool summers in our record during the seventeenth century. Besides 1601, each of the individual years 1641/2/3 (rank 3/28/10); 1666/7/8/9 (12/27/40/16); 1675 (8); 1659(6) and 1698/9(9/11) have low mean density ($<-2.0\sigma$; see Table 1) and are all among the lowest 5% of values. This implies that at least six large climatically effective eruptions occurred that century, perhaps even eight, if one considers that two of these groups probably represent double events. The early 1640s seems to be such a group, with the large 1641 signal particularly prominent (Fig. 2). Multiple eruptions are suggested by some of the better dated Greenland²⁴ and Antarctic⁷ ice cores at this time, and even though they are only datable to within ± 2 years at best, a number

indicate a relatively large volcanic signal at about 1641/2, sometimes attributed to either or both of Komaga-Take, Japan (1640) or Awu, Indonesia (1641). While mindful of not assuming a necessary link between VEI and climate, we note that the ascribed magnitudes of these eruptions (VEI 5 and 5?)¹⁴ seem too low to account for the extreme value for 1641 in NDH1 (-4.3σ) and it is, therefore, interesting to note recent work that provides convincing evidence, not only of a long-standing misidentification of the Awu eruption (the volcano was in fact Mt Parker, Mindanao) but that this eruption was probably of a magnitude equivalent to the 1991 VEI 6 eruption of Pinatubo, Philippines (C. Newhall, personal communication, and J. P. Witter and S. Self, manuscript in preparation). (A VEI value followed by “?” indicates uncertainty in the original magnitude attribution; see Table 2 legend.)

The small dating uncertainty, even in the best dated ice cores, allows us to suggest a tentative alignment between several of the other seventeenth century extremes in our record: in 1666, 1695 and 1698, with acid layers in both Greenland^{15,16} and Antarctica⁷, and in 1675 with the best dated Antarctic core⁷. This implies that at least some of these dates may indicate global scale events. However, other than Gamkonora, Indonesia, in 1673 (VEI 5?) the only possible cause for any of these extremes, listed in Table 1, is the large eruption of Long Island, New Guinea (VEI 6). The current best estimates of its date are either in the range 1630–1670 based on radiocarbon³³, and near 1690 based on Pb²¹⁰ (ref. 34). Historical evidence of revegetation of the area by the end of the seventeenth century³³ suggests an earlier date in this range, allowing speculation that either our 1666 (rank 12) or 1675 (8) events are possible evidence of this eruption. Either way, our data (and to a lesser extent the ice acidity data) suggest at least three significant eruptions in the late seventeenth century that remain otherwise unrecognized. According to their frequency of occurrence in Table 1, more significant climatically effective eruptions occurred during the seventeenth

century (11) than in any of the other five centuries of our record (for which frequencies ranged between 1 and 7).

The only other known large eruption in the past 600 years (VEI 6) to which we have not yet made reference is Billy Mitchell, southwest Pacific. This is listed¹⁴ with an uncorrected radiocarbon age of 1580 ± 20 , which, when corrected, could give a date anywhere in the range 1480–1640, though with more likelihood of a date around 1500³⁵. There are two candidates for the corresponding signal in our record: 1495 (rank 23) and 1587 (17). Both have apparent equivalents in at least one Antarctic core¹⁷ and 1587 is listed as an acid event in one Greenland ice core (though an earlier ~1570 event is ascribed to Billy Mitchell)¹⁶.

Besides large individual eruptions, our data show how closely spaced multiple eruptions can reduce hemisphere temperatures on decadal and multi-decadal timescales, as occurred in the mid-fifteenth century, the 1640s, 1660s and 70s, 1690s and the 1810s. The apparent cluster of possible eruptions in the seventeenth century, suggested by our data, perhaps in combination with lower solar irradiance, may have been a contributing factor in the extended hemispheric cooling that occurred at that time³⁶. □

Methods

Each of the 383 individual 'site' chronologies (Fig. 2) is the average of absolutely dated time-series of annual maximum latewood density measurements from multiple cores sampled from within different trees. Each constituent series is first transformed into dimensionless indices by taking residuals from a generalized exponential fit through the raw data. Each site chronology was normalized so that it had zero mean and unit standard deviation over AD 1901–40 (see ref. 2 for more details). All chronologies cover at least the period 1891–1973 but many are much longer (for example, there are 287 back to 1800, 159 to 1700, 75 to 1600 and 8 back to 1400).

Two hemispheric mean timeseries were constructed from these data, covering the period 1400–1994. NHD1 was formed as the average of 8 prior regional average series (the regions are shown in Fig. 2), while NHD2 is the direct average of all 383 chronologies. When averaging, the sample size (n) is time-dependent and increased variance in parts of the average series would normally arise because of a diminishing number of constituent chronologies. The effect was corrected for in all three averaging stages (that is, to regions, to NHD1 or to NHD2) by scaling the mean series by the square root of the effective number (n') of independent samples available (refs. 2, 37) where $n' = n/[1 + (n - 1)\bar{r}]$; here $\bar{r}$ is the mean interseries correlation between the n available samples, a measure of the common growth forcing 'signal'.

Series NHD1 is preferred, because the two-stage averaging reduces any bias that NHD2 might have towards a region with a large number of chronologies. Nevertheless, it would be incorrect to average the regional time-series together on an equal basis when one series is based on very few chronologies, so the regional series were first weighted by their expressed population signal, EPS (ref. 38). EPS measures the extent to which an average time-series of a finite number of sample series approximates the theoretical population average (that is, infinitely replicated) time-series ($\text{EPS} = n'/\bar{r}$). If the time-dependent sample is very small, the time-dependent EPS will be low and the region will be given a lower weighting. Both NHD1 and NHD2 were then normalized with respect to their means and standard deviations over the 1881–1960 period.

To derive the correlations between the density series and large-scale temperature averages cited in the text, we compared the NHD1 series with two instrumental temperature time-series, both the average of summer monthly means for April–September (ref. 9) over the period 1881–1960. We note that the period after 1960 was not used, as stated in the main text, because of concern over a recent possibly anthropogenic decline in hemispheric-average tree-ring density (ref. 2). One of the temperature series was the average only of NH grid-box series corresponding to the locations of the chronologies (NHCT), again variance corrected for reduced data in the early record. The other temperature series is the published average of all available data (land and marine) in the NH (NHLMT)⁹.

To assess the likely strength of temperature response in the earlier sections of our density series, where the sample of chronologies is considerably smaller, we rebuilt NHD1 and NHD2 but using a number of 'fixed grids' based on the

specific samples available back to certain years. The NHD2 series, constructed using just the 287 chronologies that extend back to 1800, correlates with NHCT (over 1881–1960) at 0.73 and with NHLMT at 0.55. Using the reduced samples extending back to at least 1700, 1600 and 1400 yields correlations of 0.63, 0.44 and 0.42 against NHCT, and 0.46, 0.35 and 0.32 with NHLMT. Correlations for NHD1 are as good or better and we expect that the relationships for extreme values would be stronger.

We have produced approximate estimates of NH mean summer temperature anomalies (Table 1 and Fig. 1) based on our NHD1 series by regressing it against NHLMT (adjusted to be anomalies from the 1881–1960 mean). This provides temperature anomalies with 95% confidence limits of $\sim \pm 0.3^\circ\text{C}$.

Received 13 October 1997; accepted 31 March 1998.

- Schweingruber, F. H. & Briffa, K. R. in *Climate Variations and Forcing Mechanisms of the Last 2000 Years* (eds Jones, P. D., Bradley, R. S. & Jouzel, J.) 43–66 (Springer, Berlin, 1996).
- Briffa, K. R. *et al.* Reduced sensitivity of recent tree-growth to temperature at high northern latitudes. *Nature* **391**, 678–682 (1998).
- Briffa, K. R., Jones, P. D. & Schweingruber, F. H. Summer temperatures across northern North America: regional reconstructions from 1760 using tree-ring densities. *J. Geophys. Res.* **99**(D12), 25835–25844 (1994).
- Briffa, K. R., Jones, P. D. & Schweingruber, F. H. Tree-ring reconstructions of summer temperature patterns across western North America since 1600. *J. Clim.* **5**, 735–754 (1992).
- Jones, P. D., Briffa, K. R. & Schweingruber, F. H. Tree-ring evidence of the widespread effects of explosive volcanic eruptions. *Geophys. Res. Lett.* **22**, 1333–1336 (1995).
- Bradley, R. S. & Jones, P. D. in *Climate Since A.D. 1500* (eds Bradley, R. S. & Jones, P. D.) 606–622 (Routledge, London, 1992).
- Cole-Dai, J., Mosley-Thompson, E. & Thompson, L. G. Annually resolved southern hemisphere volcanic history from two Antarctic ice cores. *J. Geophys. Res.* **102**, 16761–16771 (1997).
- Crowley, T. J., Cristie, T. A. & Smith, N. R. Reassessment of Crete (Greenland) ice core acidity/volcanism link to climate change. *Geophys. Res. Lett.* **20**, 209–212 (1993).
- Jones, P. D. & Briffa, K. R. Global surface air temperature variations during the twentieth century: Part 1, spatial, temporal and seasonal details. *Holocene* **2**, 165–179 (1992).
- Lamb, H. H. Volcanic dust in the atmosphere; with a chronology and assessment of its meteorological significance. *Phil. Trans. R. Soc. Lond. A* **266**, 425–553 (1970).
- Newhall, G. C. & Self, S. The Volcanic Explosivity Index (VEI): an estimate of explosive magnitude for historical volcanism. *J. Geophys. Res.* **87**, 1231–1238 (1982).
- Robock, A. & Free, M. P. in *Climate Variations and Forcing Mechanisms of the Last 2000 Years* (eds Jones, P. D., Bradley, R. S. & Jouzel, J.) 533–546 (Springer, Berlin, 1996).
- Harington, C. R. (ed.) *The Year Without a Summer?* (Canadian Museum of Nature, Ottawa, 1992).
- Simkin, T. & Siebert, L. *Volcanoes of the World* 2nd edn (Geoscience, Tucson, 1994).
- Hammer, C. U., Clausen, H. B. & Dansgaard, W. Greenland ice sheet evidence of post-glacial volcanism and its climatic impact. *Nature* **288**, 230–235 (1980).
- Zielinski, G. A. *et al.* Record of volcanism since 7000 B.C. from the GISP2 Greenland ice core and implications for the volcano-climate system. *Science* **264**, 948–952 (1994).
- Moore, J. C., Narita, H. & Maeno, N. A continuous 770-year record of volcanic activity from East Antarctica. *J. Geophys. Res.* **96**, 17353–17359 (1991).
- Delmas, R. J., Kirchner, S., Palais, J. M. & Petit, J. R. 1000 years of explosive volcanism recorded at the South Pole. *Tellus B* **44**, 335–350 (1992).
- Langway, C. C., Jr, Osada, K., Clausen, H. B., Hammer, C. U. & Shoji, H. A 10-century comparison of prominent volcanic events in ice cores. *J. Geophys. Res.* **100**, 16241–16247 (1995).
- Morgan, V. I. *et al.* Site information and initial results from deep ice drilling on Law Dome, Antarctica. *J. Glaciol.* **43**, 3–10 (1997).
- Dai, J. E., Mosley-Thompson, E. & Thompson, L. G. Ice core evidence for an explosive tropical eruption 6 years preceding Tambora. *J. Geophys. Res.* **96**, 17361–17366 (1991).
- Self, S., Rampino, M. R. & Carr, M. J. A reappraisal of the 1835 eruption of Cosigüina and its atmospheric impact. *Bull. Volcanol.* **52**, 57–65 (1989).
- Scuderi, L. C. Tree-ring evidence for climatically effective volcanic eruptions. *Quat. Res.* **34**, 67–85 (1990).
- Zielinski, G. A. Stratospheric loading and optical depth estimates of explosive volcanism over the last 2100 years derived from the Greenland ice sheet Project 2 ice core. *J. Geophys. Res.* **100**, 20937–20955 (1995).
- Palais, J. M., Kirchner, S. & Delmas, R. J. Identification of some global volcanic horizons by major elemental analysis of fine ash in Antarctic ice. *Ann. Glaciol.* **14**, 216–220 (1990).
- Thompson, L. G., Mosley-Thompson, E., Dansgaard, W. & Grootes, P. M. The Little Ice Age as recorded in the stratigraphy of the tropical Quelccaya Ice Cap. *Science* **234**, 361–364 (1986).
- de Silva, S. L. & Zielinski, G. A. Global influence of the AD 1600 eruption of Huaynaputina, Peru. *Nature* **393**, 455–458 (1998).
- Robin, C., Monzier, M. & Eissen, J.-P. Formation of the mid-fifteenth century Kuwae caldera (Vanuatu) by an initial hydroelastic and subsequent ignimbritic eruption. *Bull. Volcanol.* **56**, 170–183 (1994).
- Monzier, M., Robin, C. & Eissen, J.-P. Kuwae (1425 AD): the forgotten caldera. *J. Volcanol. Geotherm. Res.* **59**, 207–218 (1993).
- Pang, K. D. Climate impact of the mid-fifteenth century Kuwae caldera formation, as reconstructed from historical and proxy data. *Eos* (Fall Suppl.) **74**(43), 106 (1993).
- LaMarche, V. C. & Hirschboeck, K. K. Frost rings in trees as records of major volcanic eruptions. *Nature* **307**, 121–126 (1984).
- Delfin, F. G. Jr *et al.* Geological, ¹⁴C, and historical evidence for a 17th century eruption of Parker volcano, Mindanao, Philippines. *J. Geol. Soc. Philippines* **52**, 25–42 (1997).
- Blong, R. J. *The Time of Darkness* (Australian Nat. Univ. Press, Canberra, 1982).
- Oldfield, F., Appleby, P. G. & Thompson, R. Palaeoecological studies of lakes in the highlands of Papua New Guinea. I. The chronology of sedimentation. *J. Ecol.* **68**, 457–477 (1980).
- Stuiver, M. & Becker, B. High-precision decadal calibration of the radiocarbon time scale, A.D. 1950–6000 B.C. *Radiocarbon* **35**, 35–65 (1993).
- Lean, J., Beer, J. & Bradley, R. Reconstruction of solar irradiance since 1610: Implications for climatic change. *Geophys. Res. Lett.* **22**, 3195–3198 (1995).
- Osborn, T. J., Briffa, K. R. & Jones, P. D. Adjusting variance for sample-size in tree-ring chronologies and other regional-mean timeseries. *Dendrochronologia* **15**, (in the press).
- Briffa, K. R. in *Analysis of Climatic Variability: Applications of Statistical Techniques* (eds von Storch, H. & Navarra, A.) 77–94 (Springer, Berlin, 1995).

Supplementary Information is available on Nature's World-Wide Web site (<http://www.nature.com>) or as paper copy from the London editorial office of Nature.

Acknowledgements. The current development of the Northern Hemisphere densitometric network is funded by the Swiss National Science Foundation (F.H.S.) and the European Community under ADVANCE-10K (K.R.B.). P.D.J. is supported by the US Department of Energy, and T.J.O. by NERC. We thank E. Cook, T. Crowley, M. Free, V. Morgan, E. Mosley-Thompson, C. Newhall, F. Oldfield, A. Robock and S. Self for advice and data.

Correspondence and requests for materials should be addressed to K.R.B. (e-mail: k.briffa@uea.ac.uk).

Global influence of the AD1600 eruption of Huaynaputina, Peru

Shanaka L. de Silva* & Gregory A. Zielinski†

* Department of Geography, Geology, and Anthropology, Indiana State University, Terre Haute, Indiana 47809, USA

† Climate Change Research Center, University of New Hampshire, Durham, New Hampshire 03824-3525, USA

It has long been established that gas and fine ash from large equatorial explosive eruptions can spread globally, and that the sulphuric acid that is consequently produced in the stratosphere can cause a small, but statistically significant, cooling of global temperatures^{1,2}. Central to revealing the ancient volcano–climate connection have been studies linking single eruptions to features of climate-proxy records such as found in ice-core^{3–5} and tree-ring^{6–8} chronologies. Such records also suggest that the known inventory of eruptions is incomplete, and that the climatic significance of unreported or poorly understood eruptions remains to be revealed. The AD1600 eruption of Huaynaputina, in southern Peru, has been speculated to be one of the largest eruptions of the past 500 years; acidity spikes from Greenland and Antarctica ice^{3–5}, tree-ring chronologies^{6–8}, along with records of atmospheric perturbations in early seventeenth-century Europe and China^{9,10}, implicate an eruption of similar or greater magnitude than that of Krakatau in 1883. Here we use tephra deposits to estimate the volume of the AD1600 Huaynaputina eruption, revealing that it was indeed one of the largest eruptions in historic times. The chemical characteristics of the glass from juvenile tephra allow a firm cause–effect link to be established with glass from the Antarctic ice, and thus improve on estimates of the stratospheric loading of the eruption.

Volcan Huaynaputina, located at 16° 35' S, 70° 52' W in the Moquegua region of southern Peru (Fig. 1), consists of three craters situated at 4,500 m within the amphitheatre of an ancient strato-volcano. A compilation of historical and parochial literature¹¹ reveals that the volcano formed during an eruption that began on 19 February 1600 and continued until 5 March. The fall deposit covered an area of at least 300,000 km² in southern and west central Peru, western Bolivia and north Chile (Fig. 1). Ashfall was reported on the major cities of Lima, La Paz (Bolivia) and Arica (Chile) as well as on a ship 1,000 km to the west¹¹. The fall seems to have been

distributed predominantly to the west and north. Local proximal pyroclastic flows and secondary mass flows were insignificant in volume. This large plinian eruption destroyed local communities with a loss of over 1,000 lives and caused considerable damage to the major cities of Arequipa and Moquegua. The loss of farmland, crops, livestock, vineyards, and water resources compounded the significant economic burden on this region. The whole socio-economic infrastructure of a large part of Peru, and maybe that of

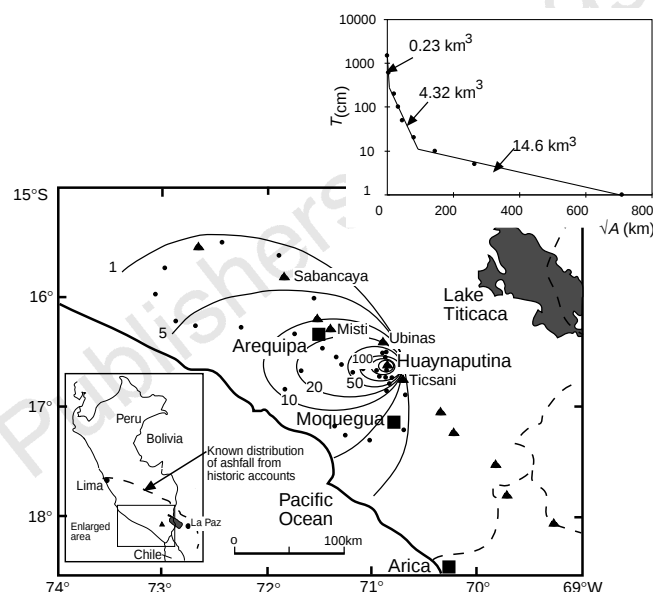


Figure 1 Isopach map of Huaynaputina tephra deposit showing the strong westerly distribution of the deposit. Black triangles are volcanoes of the modern volcanic arc and the most recently active are labelled; small dots are data points. Volcan Ticsani, the site of a similar but smaller eruption to Huaynaputina, is also shown. Values associated with contours are in centimetres; two innermost contours not labelled are 200 and 600 cm respectively, and the 1,500 cm contour is covered by the symbol for the volcano. The graph in the upper right corner shows a plot of thickness (T) on a logarithmic scale against the square root of isopach area^{31,32}. The thickness data are described by three line segments that define a typical exponential profile. Integration beneath the line segments with respect to area, following the method of Fierstein and Nathenson³², and removing the volume achieved by extrapolating the data to zero thickness³¹, yields the volumes noted for each segment and the minimum volume of tephra of 19.2 km³. Negative error in this estimate is limited to measurement and area estimation errors and is ~10% for this data set. Positive errors far outweigh this, as discussed in the text. Inset (bottom left) shows the known distribution of the fall from the AD1600 eruption (broken line) based on a compilation of parochial and historical accounts¹¹. The box in the inset shows the area covered in the larger map. Note that our data cover only a small portion of the known distribution of the AD1600 tephra fall, albeit along the main axis. As most of the distal tephra has been lost, a considerable volume of distal tephra cannot be measured.

Table 1 Main oxide composition of individual juvenile clasts and volcanic glass from the Huaynaputina eruption compared with tephra from the early 1600s of the Antarctica and Greenland ice cores

	Huaynaputina bulk rock* (21 samples)	Huaynaputina glass* ($n = 103$; 6 samples)	Huaynaputina glass† ($n = 7$)	South Pole glass† ($n = 16$)	GISP2 glass* ($n = 16$)
SiO ₂	65.16 ± 0.70	72.82 ± 0.68	75.76 ± 0.53	72.92 ± 2.04	62.1 ± 4.7
TiO ₂	0.58 ± 0.03	0.26 ± 0.20	0.27 ± 0.06	0.34 ± 0.20	1.1 ± 0.5
Al ₂ O ₃	16.68 ± 0.40	15.20 ± 1.21	13.85 ± 0.15	13.98 ± 0.91	16.2 ± 2.5
FeO	3.97 ± 0.17	1.31 ± 0.25	1.10 ± 0.14	1.37 ± 0.28	8.2 ± 3.5
MgO	1.78 ± 0.12	0.30 ± 0.13	0.27 ± 0.14	0.44 ± 0.20	1.9 ± 1.0
CaO	3.88 ± 0.17	1.61 ± 0.32	1.22 ± 0.07	1.41 ± 0.35	3.9 ± 1.7
Na ₂ O	4.57 ± 0.09	4.53 ± 0.34	3.43 ± 0.24	4.12 ± 0.73	4.0 ± 1.1
K ₂ O	2.79 ± 0.15	3.82 ± 0.18	3.89 ± 0.15	3.76 ± 0.28	2.5 ± 0.9

Errors are 1 σ . n = number of individual analyses. Bulk rock analyses are of juvenile clasts from proximal and medial locations.

* This study.

† From ref. 16.

Table 2 Climate proxy data and atmospheric phenomena indicative of explosive volcanism in the earliest 1600s with known eruptions AD 1600–04

Ice core data					
Location	Signal type/characteristics	Year of signal with error	Total volcanic flux kg km ⁻² (time)	Stratospheric mass loading (Mt)	
				Huaynaputina	Tambora
Greenland					
GISP2 ⁴	SO ₄ ²⁻ doublet	1603–04 (±2)	36 (2 yr)	87	86
GRIP ²¹	ECM doublet	1601–02 (±1)	36 (2.3 yr)	52	38
	SO ₄ ²⁻ doublet	1601–02 (±1)	54 (2 yr)	78	101
Crete³					
	ECM doublet	1601–02 (±1)	61 (2 yr)	50	150
Antarctica					
South Pole ⁵	SO ₄ ²⁻ peak	1601 (±5)	22.5 (3–4 yr)	100	307
Siple Station ²²	SO ₄ ²⁻ peak	1599 (±2)	34 (~4 yr)	n.d.	n.d.
Dyer Plateau ²²	SO ₄ ²⁻ peak	1599 (±2)	30 (~4 yr)	n.d.	n.d.

Tree ring

Phenomenon	Location	Year	Comment
Narrow ring widths	Fennoscandia	1601	Calibrated to fourth coldest summer over past 1,500 years ²³
	Western USA	1601	Calibrated to coldest summer in most of western North America over past 400 years ²⁴
	Northern Hemisphere	1601	Calibrated to coldest summer over past 600 years ⁷⁸
	Sierra Nevada	1601	Cold summer ²⁵
Frost rings	Western USA	1601	Below freezing temperatures during growing season ⁶
Light rings	Upper tree line, Quebec	1601 and 1603	Cold temperatures near end of growing season ²⁶

Atmospheric phenomena

Location	Comments
Scandinavia	Sun dimmed by constant haze in 1601 (ref. 9)
Central Europe	All through 1601 until the end of July 1602, Sun and Moon “reddish, faint and lacked brilliance”. These phenomena led to minimum total (1601–04) dust veil index (DVI) estimate of 1,000 by comparison with phenomena associated with 1883 Krakatau and 1783 Laki eruptions (ref. 9).
Iceland	Summer of 1601, Sun pale and sunlight so faint that shadows not cast. Sky appeared cloudy and pale even when no clouds present. Winter of 1601–02 extremely severe (G. Larsen, personal communication, after ref. 27).
China	1603, red and dim Suns; 1604, large sunspots (ref. 10).
Graz, Austria	Darkened partial lunar eclipse in December 1601 implies a Northern Hemisphere average optical depth of 0.10, comparable to El Chichón and Agung, although if the dimness of the December 1601 partial eclipse was from Huaynaputina aerosols, then the 0.10 optical depth would not have been the maximum optical depth, but an optical depth following almost two years of aerosol loss from the stratosphere (R. Keen, personal communication after refs 28, 29).

Potential source eruptions (ref. 30 unless stated otherwise)

Volcano (latitude) tephra	Date of eruption	VEI	Volume (km ³)
Asama (Japan, 36° N)	January 1600	3?	?
Huaynaputina (Peru, 17° S) (this work)	February 1600	6	>19
Suwanose-jima (Ryuku Is., 30° N)	1600 (year questionable)	4+	?
Iwaki (Japan, 41° N)	February 1604	3?	?

n.d., not determined, but considering that flux values are slightly higher than the South Pole values, loading estimates would be similar to or slightly above the 100 Mt estimate.

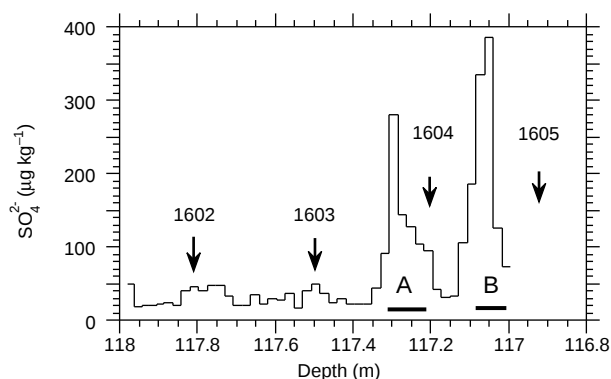


Figure 2 Subseasonal record of sulphate concentrations in the GISP2 ice core for the period AD 1601–05. Arrows denote mid-summer for the year shown. Maximum dating error for this period is only ~2 years. It is believed, from correlation with the other ice cores, that the ages shown might be 1 or 2 years too young, although the red Suns observed in China in 1603 (Table 2) indicate that there still was a significant stratospheric aerosol loading in that year. Deposition of those aerosols on Greenland could easily have been in 1604. Even if the two sulphate spikes correspond to 1601/1602 and 1602/1603, these are ages within the error of the GISP2 dating shown on the figure. Flux of the volcanic sulphate component of the record is 13 kg km⁻² per 0.5 year for peak A and 21 kg km⁻² per 0.5 yr for peak B from the detailed (~14 samples yr⁻¹) results presented here. Total volcanic flux of 34 kg km⁻² per 2 yr agrees well with the 36 kg km⁻² per 2 yr obtained from the original biyearly sampling done on the GISP2 core⁴ and given in Table 1. Total volcanic sulphate flux is determined by the sum of the products of concentration, measured ice density and length for each sample. Volcanic sulphate concentration is the total concentration minus average background levels. Sea-salt sulphate was removed from the calculations. A and B also indicate the length of sections of ice filtered in an attempt to locate and identify volcanic glass with a scanning electron microscope and electron microprobe. Figure modified from ref. 4.

Bolivia and Chile, was devastated by the eruption. A full recovery of the economy took almost 150 years (ref. 11).

We calculated new estimates of the volume erupted on the basis of field measurements of present-day tephra thicknesses and the construction of an isopach map (Fig. 1). From this a minimum erupted tephra volume of $\sim 19.2 \text{ km}^3$ was calculated. Preliminary measurements of bulk tephra densities yield a weighted average density of $\sim 1,200 \text{ kg m}^{-3}$, which yields a minimum dense rock equivalent (DRE) or magma equivalent volume of $\sim 9.6 \text{ km}^3$. This is the first and only volume estimate based on field measurements for this eruption. The true volume of the eruption might be considerably larger. First, our estimate does not include distal tephra outside the mapped isopachs. Second, our measurements cover only a portion of the area of the fall deposit, albeit along the main axis of the deposit (Fig. 1). Last, historical and anecdotal information¹¹ clearly report greater thicknesses of tephra deposition than we find in the field today. For instance, during the first 24 h alone, over 20 cm of sand-sized tephra fell in Arequipa. Over 1 m of ash was reported in regions to the north and west of Arequipa after the eruption; today we find a maximum of 10 cm. Even allowing for compaction of distal ash, this still suggests that a considerable amount of the tephra has been lost, most obviously to deflation, mass wasting, and agriculture. Speculation about true volume aside, our minimum estimate for the tephra deposit of 19.2 km^3 firmly establishes this eruption as a magnitude 6 on the Volcanic Explosivity Index (VEI (ref. 12)). This alone places this eruption firmly in the realms of the largest historic eruptions such as Krakatau in 1883 (ref. 13) (18 km^3 tephra; VEI ~ 6), Novarupta (Katmai) in 1912 (ref. 14) ($\sim 20\text{--}25 \text{ km}^3$ tephra; VEI ~ 6) and Mt Pinatubo in 1991 (ref. 15) (10 km^3 tephra; VEI 6).

New geochemical data on the products of the eruption now allow us to establish a much firmer link between this eruption and the volcanic glass found in the ice-core record. These data reveal that the magma of the AD1600 eruption was a monotonous calc-alkaline dacite (Table 1). The juvenile clasts of non-expanded to moderately expanded dacite contain $\sim 30\text{--}35 \text{ vol.}\%$ phenocrysts of plagioclase > amphibole > oxides > apatite set in a matrix of fresh glass. Huaynaputina magmatic compositions are distinct from other late Cenozoic to recent volcanics of similar composition in this region; of particular note here is that the ratio of Na_2O to K_2O is greater than 1 for the Huaynaputina magma (Table 1). Our data for the glass reveal this same distinctive character (Table 1) and are a much closer match to the analyses of volcanic glass in the South Pole ice core than presented in previous studies¹⁶. These data provide the strongest support yet for the presence of volcanic glass from Huaynaputina in the South Pole ice core and thus the climatic impact of the eruption might be better delineated.

Climate-proxy data (Table 2) indicate that the earliest part of the seventeenth century and especially the summer of 1601 was the coldest of the past 600 years throughout much of the Northern Hemisphere and among the coldest of the past 1,500 years in Fennoscandia. High levels of H_2SO_4 found in ice-core layers dated to 1599–1604 (with dating error) from both polar regions, together with observed atmospheric phenomena (Table 2), strongly suggest that volcanism, and presumably an equatorial eruption such as Huaynaputina, had some role in forcing that cooling¹⁷. In fact, the mean estimate of the stratospheric loading of H_2SO_4 based on the Greenland ice core record is $67 \pm 16 \text{ Mt}$ or twice that estimated for the Mt Pinatubo eruption in 1991 ($\sim 30 \text{ Mt}$; ref. 18). The 67 Mt estimate is $\sim 78\%$ of the stratospheric loading estimated for the 1815 Tambora eruption from data from these same ice cores (Table 2). The only estimate for Huaynaputina from Antarctica is 100 Mt H_2SO_4 (Table 2). These values are substantial enough to force climate cooling.

How much of this stratospheric loading and subsequent cooling can confidently be attributed to Huaynaputina? On the basis of the strong evidence for Huaynaputina glass in the South Pole core, the

100 Mt estimate can reasonably be related entirely to the Huaynaputina eruption.

The Greenland data are more problematic because the volcanic signal occurs as a distinct doublet peak in Northern-Hemisphere ice cores, and we cannot link the earlier peak of the GISP2 core doublet (Fig. 2) to Huaynaputina because the glass compositions do not match (Table 1). This implicates another eruption as the main source of the sulphate; glass from Asama, Japan¹⁹, do not match either, so the most likely source is either the Suwanose-Jima eruption (Table 2) or a previously undocumented eruption originating in the Northern Hemisphere. Higher total volcanic flux values in signals from the early 1600s in the Greenland ice cores relative to Antarctica, particularly given the 17°S location of Huaynaputina, would probably result from an eruption directly upwind from Greenland. This notwithstanding, it is reasonable to argue that ash particles are less likely than aerosols to be transported interhemispherically from a tropical eruption such as Huaynaputina: the lack of ash particles does not categorically negate a role for this eruption. We therefore make the conservative assumption that the younger sulphate spike in the GISP2 core (B in Fig. 2) is from Huaynaputina. This yields a volcanic sulphate flux of 21 kg km^{-2} per 0.5 yr from the GISP2 core, which translates to a stratospheric mass loading estimate of $\sim 50 \text{ Mt}$ (ref. 4) or two-thirds of the total GISP2 volcanic signal over these two years. If a similar proportion of Huaynaputina acid is found in the total signals for the other Greenland cores, then the mean Northern Hemisphere loading for Huaynaputina would be $\sim 42 \pm 10 \text{ Mt}$ or almost 50% greater loading than Pinatubo, but only one-half of the Tambora loading for the Northern Hemisphere. A composite estimate of stratospheric loading from Greenland snow records for the equatorial El Chichón eruption agreed well with satellite estimates²⁰; thus our estimate of the stratospheric loading for the Huaynaputina eruption derived from this series of Greenland ice cores should be reliable. The 100 Mt loading estimate for the Southern Hemisphere, based on the South Pole core (Table 2), might be a reflection of the southern equatorial location of the volcano and the greater distribution of aerosols to southern polar regions. These two hemispheric estimates yield a global average of $\sim 70 \text{ Mt H}_2\text{SO}_4$ for the Huaynaputina eruption. This is only 35% of the $\sim 200 \text{ Mt}$ (global average) estimated for Tambora, the greatest sulphur-producing eruption of at least the past ~ 750 years. Nevertheless, these data indicate that the Huaynaputina eruption was very capable of forcing global climate, and its atmospheric perturbation was not only far greater than any eruption in the twentieth century, it is also one of the greatest of historical time. □

Received 7 July 1997; accepted 10 April 1998.

1. Cadle, R. D., Kiang, C. S. & Louis, J. F. The global scale dispersion of the eruption clouds from major volcanic eruptions. *J. Geophys. Res.* **81**, 3125–3132 (1976).
2. Mass, C. F. & Portman, D. A. Major volcanic eruptions and climate: a critical evaluation. *J. Clim.* **2**, 566–593 (1989).
3. Hammer, C. U., Clausen, H. B. & Dansgaard, W. Greenland ice sheet evidence of post-glacial volcanism and its climatic impact. *Nature* **288**, 230–235 (1980).
4. Zielinski, G. A. Stratospheric loading and optical depth estimates of explosive volcanism over the last 2100 years derived from the GISP2 Greenland ice core. *J. Geophys. Res.* **100**, 20937–20955 (1995).
5. Delmas, R. J., Kirchner, S., Palais, J. M. & Petit, J. R. 1000 years of explosive volcanism recorded at the South Pole. *Tellus* **B44**, 335–350 (1992).
6. LaMarche, V. C. & Hirschboeck, K. K. Frost rings in trees as records of major volcanic eruptions. *Nature* **307**, 121–126 (1984).
7. Jones, P. D., Briffa, K. R. & Schweingruber, F. H. Tree-ring evidence of the widespread effects of explosive volcanic eruptions. *Geophys. Res. Lett.* **22**, 1333–1336 (1995).
8. Briffa, K. R., Jones, P. D., Schweingruber, F. H. & Osborn, T. J. Influence of volcanic eruptions on Northern Hemisphere summer temperatures over 600 years. *Nature* **393**, 450–455 (1998).
9. Lamb, H. H. Volcanic dust in the atmosphere; with a chronology and assessment of its meteorological significance. *Phil. Trans. R. Soc. Lond. A* **266**, 425–533 (1970).
10. Scuderi, L. A. Oriental sunspot observations and volcanism. *Q. J. R. Astron. Soc.* **31**, 109–120 (1990).
11. de Silva, S. L., Alzueta, J. & Salas, G. in *Volcanic Disasters in Human Antiquity* (eds Heiken, G. & McCoy, F.) (Geol. Soc. Am. Spec. Paper, in the press).
12. Newhall, C. & Self, S. The volcanic explosivity index (VEI): an estimate of explosive magnitude for historical volcanism. *J. Geophys. Res.* **87**, 1231–1238 (1982).
13. Simkin, T. & Fiske, R. *Krakatau 1883: the Volcanic Eruption and its Effects* (Smithsonian Inst., Washington DC, 1983).
14. Hildreth, W. New perspectives on the eruption of 1912 Valley of Ten Thousand Smokes, Katmai National Park, Alaska. *Bull. Volcanol.* **49**, 680–693 (1987).
15. Scott, W. E. et al. in *Fire Mud: Eruptions and Lahars of Mount Pinatubo, Philippines* (eds Newhall, C. G.

- & Punongbayan, R. S.) 545–571 (Univ. Washington Press, Seattle, 1996).
16. Palais, J. M., Kirchner, S. & Delmas, R. J. Identification of some global volcanic horizons by major element analysis of fine ash in Antarctic ice. *Ann. Glaciol.* **14**, 216–220 (1990).
 17. Pyle, D. M. On the 'climate effectiveness' of volcanic eruptions. *Quat. Res.* **37**, 125–129 (1992).
 18. McCormick, M. P., Thomason, L. W. & Trepte, C. R. Atmospheric effects of the Mt Pinatubo eruption. *Nature* **373**, 399–404 (1995).
 19. Zielinski, G. A. et al. Climatic impact of the AD 1783 eruption of Asama (Japan) was minimal: evidence from the GISP2 ice core. *Geophys. Res. Lett.* **21**, 2365–2368 (1994).
 20. Zielinski, G. A. et al. Assessment of the record of the 1982 El Chichón eruption as preserved in Greenland snow. *J. Geophys. Res.* **102**, 30031–30045 (1997).
 21. Clausen, H. B. et al. A comparison of the volcanic records over the past 4000 years from the Greenland ice core project and Dye 3 Greenland ice cores. *J. Geophys. Res.* **102**, 26707–26723 (1997).
 22. Cole-Dai, J., Mosley-Thompson, E. & Thompson, L. Annually resolved southern hemisphere volcanic history from two Antarctic ice cores. *J. Geophys. Res.* **102**, 16761–16771 (1997).
 23. Briffa, K. R. et al. Fennoscandian summers from AD 500: temperature changes on short and long timescales. *Clim. Dyn.* **7**, 111–119 (1992).
 24. Briffa, K. R., Jones, P. D. & Schweingruber, F. H. Tree-ring density reconstructions of summer temperature patterns across western North America since 1600. *J. Clim.* **5**, 735–754 (1992).
 25. Scuderi, L. A. Tree ring evidence for climatically effective volcanic eruptions. *Quat. Res.* **34**, 67–86 (1990).
 26. Filion, L., Payette, S., Gauthier, L. & Boutin, Y. Light rings in subarctic conifers as a dendrochronological tool. *Quat. Res.* **26**, 272–279 (1986).
 27. Anonymous *Annaler* **6**, 493–494 (1400–1800).
 28. Frisch, C. (ed.) *Joannis Kepleris Astronomi Opera Omnia* **2/3** (1856–1871).
 29. Keen, R. A. Volcanic aerosols and lunar eclipses. *Science* **222**, 1011–1013 (1983).
 30. Simkin, T. & Siebert, L. *Volcanoes of the World* 2nd edn (Geoscience Press, Tucson, 1994).
 31. Pyle, D. M. The thickness, volume, and grain size of tephra fall deposits. *Bull. Volcanol.* **51**, 1–15 (1989); Assessment of the minimum volume of tephra fall deposits. *J. Volcanol. Geotherm. Res.* **69**, 379–382 (1995).
 32. Fierstein, J. & Nathenson, M. Another look at the calculation of fallout tephra volume. *Bull. Volcanol.* **54**, 156–167 (1991).

Acknowledgements. We thank G. Salas, P. Francis, S. Self and the Instituto Geofísico de Peru, particularly the late M. Chang, for their collaboration on this project; P. Mayewski, L. D. Meeker, S. Whitlow and M. Twickler for their work in producing the initial sulphate time series of the GISP2 ice core; M. Germani and J. Palais for their help with the GISP2 tephra studies; J. Fierstein, M. Nathenson and N. Adams for help and discussions about the volume estimates; the members of the GISP2 community for work in developing the chronology of the core; and the Science Management Office, Polar Ice Coring Office and 109th Air National Guard for logistical support. Funding for this work has come from Indiana State University, and the National Science Foundation Petrology and Geochemistry Program, Office of Polar Programs (GISP2 work), and Atmospheric Sciences Program. The manuscript has benefited considerably from thorough reviews by D. Pyle and P. Allard.

Correspondence and requests for materials should be addressed to S.L.d.S. (e-mail: gesilva@scifac.indstate.edu).

A one-million-year-old *Homo* cranium from the Danakil (Afar) Depression of Eritrea

Ernesto Abbate*, Andrea Albanielli*, Augusto Azzaroli†, Marco Benvenuti*, Berhane Tesfamariam‡, Piero Bruni*, Nicola Cipriani*, Ronald J. Clarke§, Giovanni Ficcarelli*, Roberto Macchiarelli||, Giovanni Napoleone*, Mauro Papini*, Lorenzo Rook*, Mario Sagri*, Tewelde Medhin Tecle¶, Danilo Torre† & Igor Villa#

* Dipartimento di Scienze della Terra, Università di Firenze, 50121 Firenze, Italy

† Museo di Storia Naturale, Sezione di Geologia e Paleontologia, Università di Firenze, 50121 Firenze, Italy

‡ Eritrea National Museum, Asmara, Eritrea

§ Palaeo-Anthropology Research Unit, Department of Anatomical Sciences, University of the Witwatersrand Medical School, Johannesburg 2193, South Africa

|| Museo Nazionale Preistorico Etnografico 'L. Pigorini', Sezione di Antropologia, 00144 Roma, Italy

¶ Ministry of Energy, Water and Mineral Resources, Department of Mines, Asmara, Eritrea

Mineralogisches Institut, Abteilung für Isotopengeologie, Universität Bern, Switzerland

One of the most contentious topics in the study of human evolution is that of the time, place and mode of origin of *Homo sapiens*^{1–3}. The discovery in the Northern Danakil (Afar) Depression, Eritrea, of a well-preserved *Homo* cranium with a mixture of characters typical of *H. erectus* and *H. sapiens* contributes significantly to this debate. The cranium was found in a succession of fluvio-deltaic and lacustrine deposits and is associated with a rich

mammalian fauna of early to early-middle Pleistocene age. A magnetostratigraphic survey indicates two reversed and two normal magnetozones. The layer in which the cranium was found is near the top of the lower normal magnetozones, which is identified as the Jaramillo subchron. Consequently, the human remains can be dated at ~1 million years before present.

The cranium, two incisors and two pelvic fragments were recovered between 1995 and 1997 near the village of Buia, Eritrea, about 100 km south-south-east of Massawa in the Danakil Formation, which is more than 1,000 m thick. This unit, also known as the Red Series^{4,5}, crops out extensively in the Danakil Depression, a belt of pronounced lithospheric extension at the intersection of the Red Sea, the Gulf of Aden, and the East African rift systems⁶ (Fig. 1). It consists mainly of siliciclastic continental deposits with intercalations of lava flows and tuff beds at different levels. A Miocene to Pleistocene age is commonly assumed^{4,5,7,8}, and from a chronological and palaeo-environmental point of view, this formation can be correlated with the well-known hominid-bearing Awash Group in Ethiopia⁹.

We examined the upper 500-m-thick portion of the Danakil Formation that is composed of grey to whitish clayey silts and sands with small amounts of marls laid down in fluvio-deltaic and lacustrine environments (Fig. 2).

Ash layers containing biotite have been found in the lower portion of the section. An initial attempt to date a biotite bulk separate by ³⁹Ar/⁴⁰Ar failed because of xenocrystic contamination.

As well as freshwater gastropod and fish remains, a large amount of fossil bones were found along the entire Buia section, and five mammal-rich levels have been identified (Fig. 2). The third level from the base yielded *Homo* remains consisting of a nearly complete and undeformed adult cranium (UA 31) (Fig. 3), two pelvic fragments (UA 173), and two permanent lower incisors (UA 222 and UA 369).

In the *Homo*-bearing outcrop, silts and clays alternate with sands. The cranium and pelvic fragments come from the upper part of a layer of laminated silty clays 1.4 m thick. The UA 369 incisor was

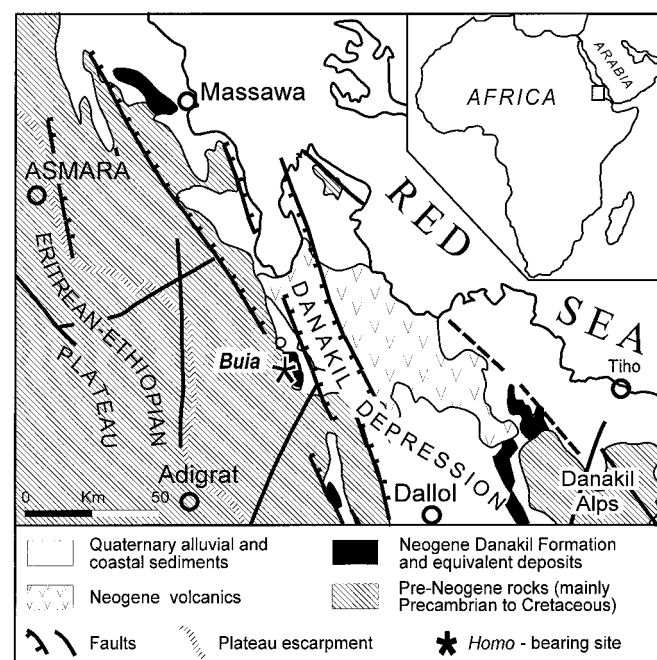


Figure 1 Simplified geology of the Northern Danakil Depression between the Eritrean-Ethiopian plateau and the Southern Red Sea. The palaeo-anthropological site near Buia is indicated by an asterisk.

found in a large sandy channel resting above this layer, and the UA 222 incisor comes from a thin lenticular sandy bed 1.5 m below the skull-bearing layer.

The *Homo*-bearing layer of the Buia section is exposed in several well-correlated outcrops in which a rich vertebrate fauna is repre-

sented by *Elephas recki*, *Hipparion* sp., *Equus* sp., *Ceratotherium simum*, Rhinocerotidae indet., *Pelorovis* sp., *Kobus ellipsiprymnus*, Bovidae indet., *Kolpochoerus* cf. *K. majus*, cf. *Kolpochoerus* sp. (advanced form), *Hippopotamus gorgops*, *Hexaprotodon karumensis*, Hyaenidae indet., Chelonia and Crocodilia. This is a typical African savanna fauna and indicates an age interval from early to early middle Pleistocene.

A magnetostratigraphic study was carried out on 97 hand samples collected in the silts and fine sands of the sequence. Characteristic directions of the palaeomagnetic vector, selected after thermal and alternative field cleaning procedures, yielded the virtual geomagnetic pole (VGP) latitudes used for defining the magnetostratigraphic succession (Fig. 2). Four magnetozones are recognized: two reversed R1 (0–115 m) and R2 (202–377 m), and two normal N1 (115–202 m) and N2 (377–500 m). The layer with the *Homo* remains is placed near the top of magnetozone N1.

On the basis of the biochronological and palaeomagnetic data, the Buia magnetozones have been correlated with the Geomagnetic Polarity Time Scale¹⁰, and the magnetozone N1 has been identified as Jaramillo. This attribution relies on the following considerations. First, the mammal fauna gives a comprehensive age interval from the early Pleistocene to the early middle Pleistocene. Second, the succession does not reveal significant gaps and can be regarded as substantially continuous. Third, the two reversed magnetozones R1 and R2 are quite thick and, presumably, cover two relatively extended periods of time. They can be safely related to the Matuyama chron. Finally, the interposed polarity normal N1 has to be identified as Jaramillo, because it is thinner than the overlying N2. If N1 were identified as the Olduvai subchron, it would depend on the assumption that different intervals had markedly different rates of deposition, and there is no evidence for this. Because the *Homo*-bearing layer is near the top of the Jaramillo subchron it can be dated to ~1 Myr before present¹⁰.

The cranium exhibits a nearly complete braincase (lacking most of the basicranium) and parts of the facial skeleton. No dental crowns remain, but the roots of the right P3, P4, M1, M2, M3 (mesial), and of the left P4, M1, and M2 are included in the alveoli.

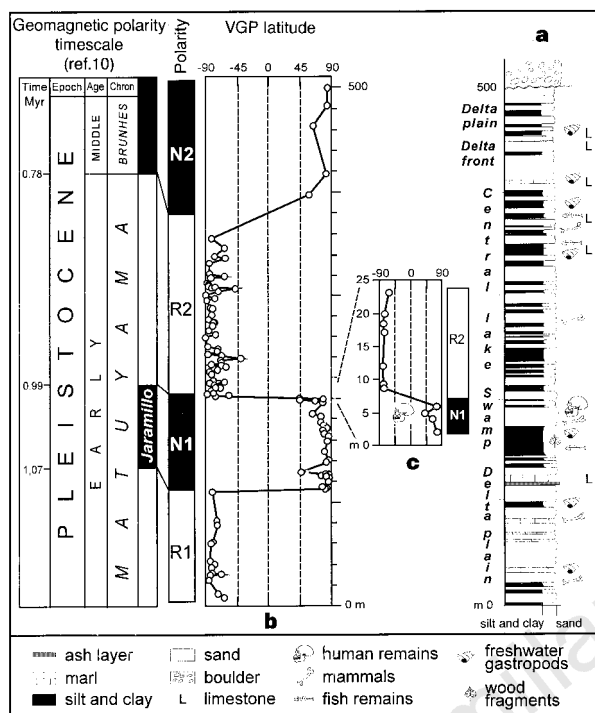


Figure 2 Stratigraphy of the Danakil Formation in the Buia area. **a**, Generalized lithostratigraphy with palaeo-environmental reconstructions. The boulder beds unconformably overlie the studied section. **b**, Composite magnetostratigraphic section obtained here and correlation with the Geomagnetic Polarity Time Scale. **c**, Detailed magnetostratigraphy of the site where the *Homo* cranium was found.

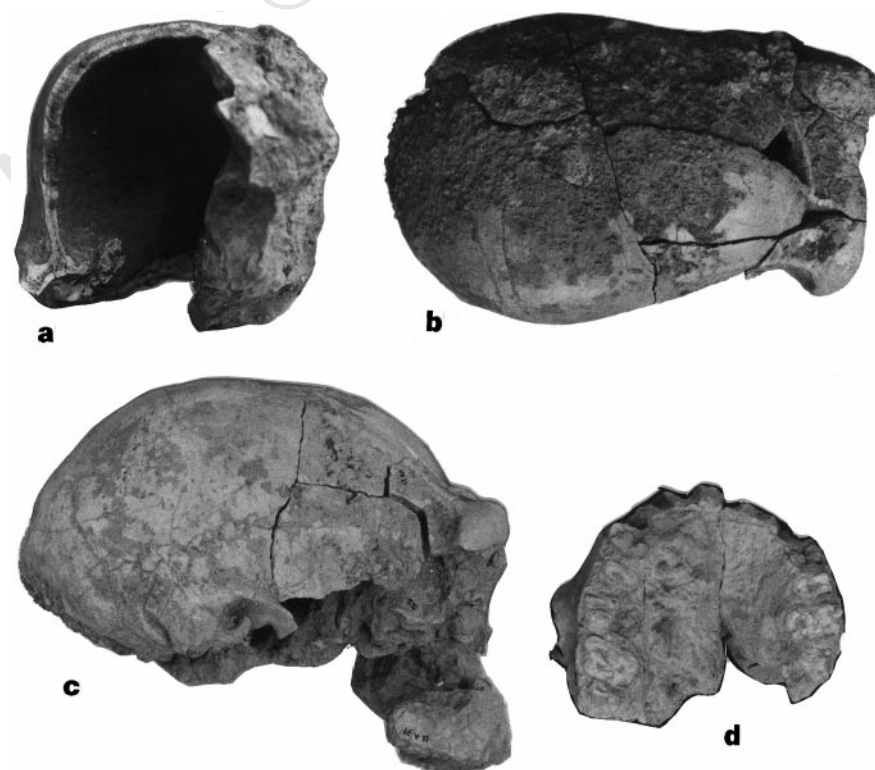


Figure 3 The UA 31 cranium. **a**, Frontal view of the braincase displaying a naturally exposed post-bregmatic coronal section through right parietal; **b**, upper view; **c**, right lateral view; **d**, basal view of the maxilla. (**a–c**, 30% of actual size; **d**, 60% of actual size).

A silty matrix still adheres to the external left half of the specimen.

For logistic reasons it has not yet been possible to clean, reconstruct and fully study the cranium, so here we offer only a preliminary assessment. For similar reasons the pelvic fragments and incisors are not described here.

A remarkable feature of UA 31 is its great anteroposterior length (204 mm) in relation to its maximum cranial breadth (130 mm). The greatest cranial breadth is located at the level of the supra-mastoid crests and with the length gives a cranial index of 63.7. Although the skull has not yet been restored and cleaned, tentative estimates suggest an endocranial capacity of 750–800 cm³.

The neurocranium is low compared to its maximum length, but high relative to its transverse diameter. The upper-middle facial skeleton is slightly concave and marked prognathism is evident in the subnasal region. The face is narrow and short. The palate is shallow, wide and short, with its length reaching less than one-third of the total cranial length as seen in basal view. The very heavily built, forwardly projecting supraorbital torus is arched and thickened medially and over the middle of each orbit (17 mm). The minimum frontal breadth is 84.0 mm. Substantial elevation of the forehead and absence of sagittal keeling are evident in frontal view.

The frontal squame is well curved to bregma and the vault outline becomes gradually flatter on the parietals. The temporal lines originate in the form of a raised crest at the posterolateral margin of the supraorbital torus, but disappear early on the parietals. A slightly thickened angular torus is present. There is no nuchal torus, but a modest external protuberance is detectable where opisthocranion coincides with inion.

The specimen also shows a modest protrusion of the mastoid–supramastoid–auricular complex lateral to the temporal squame. The mastoid process is short and broad. In coronal section, the parietal walls converge slightly inferiorly (Fig. 3); there is a high positioning of the maximum biparietal breadth (125.0 mm); the parietal thickness progressively decreases up to the midline (from 7.4 to 6.0 mm); the thickness at bregma is 6.3 mm.

The cranium from Buia shows the following combination of features, which are characteristic of the African specimens referred to *Homo erectus* and *Homo ergaster*^{11,12}: long ovoid braincase with low endocranial capacity; greatest cranial breadth across the supra-mastoid crests; massive supraorbital torus; opisthocranion coincident with inion. One trait typical of *H. sapiens* is the high position of the greatest biparietal breadth with parietal walls converging slightly inferiorly. This mosaic of primitive and progressive features increases the known morphological variation reported for early–middle Pleistocene *Homo* crania and urges caution in the specific assessment of UA 31. Until we have conducted a full comparative study, we prefer to leave open its specific allocation.

Given its chronological position, in the middle of the time interval 1.4–0.6 Myr, from which no human cranial remains of comparable integrity are known from Africa¹³, this unique Eritrean specimen provides a new perspective on the origins of modern humans. Its age indicates that morphology like that of *H. sapiens* had begun to differentiate in Africa at ~1 Myr, which is ~0.3 Myr earlier than previously estimated^{1,2,14}. □

Received 25 September 1997; accepted 26 February 1998.

- Bräuer, G., Yokoyama, Y., Falguères, C. & Mbua, E. Modern human origins backdated. *Nature* **386**, 337–338 (1997).
- Ward R. & Stringer C. Molecular handle on the Neanderthals. *Nature* **388**, 225–226 (1997).
- Bermúdez de Castro, J. M. et al. A hominid from the lower Pleistocene of Atapuerca, Spain: possible ancestor to Neanderthals and modern humans. *Science* **276**, 1392–1395 (1997).
- Bannert, D. et al. Zur Geologie der Danakil Senke (Nördliches Afar Gebiet) NE Äthiopien. *Geol. Rundsch.* **59**, 409–443 (1970).
- Brinckmann, J. & Kürsten, M. *Geological Sketchmap of the Danakil Depression* (Bundesanstalt für Bodenforschung, Hannover, 1970).
- Barberi, F. & Varet, J. Volcanism of Afar: small-scale plate tectonics implications. *Geol. Soc. Am. Bull.* **88**, 1251–1266 (1977).
- Merla, G. et al. *Geological Map of Ethiopia and Somalia 1:2,000,000 and Comment with a Map of Major Landforms* (Centro Stampa, Firenze, 1979).
- Garland, C.R. *Geological Survey of Ethiopia Memoir Vol. 1* (Berhanena Selam, Addis Ababa, 1980).
- Kalib, J. E. et al. Geology and stratigraphy of Neogene deposits, Middle Awash valley, Ethiopia. *Nature* **298**, 17–25 (1982).

- Berggren, W. A., Kent, D. V., Swisher III, C. C. & Aubry, M.-P. A revised Cenozoic geochronology and chronostratigraphy. *SEPM (special publ.)* **54**, 129–212 (1995).
- Wood, B. *Koobi Fora Research Project Vol. 4, Hominid Cranial Remains* (Clarendon, Oxford, 1991).
- Tobias, P. V. *Olduvai Gorge Vol. 4, V-IX. The Skull, Endocrasts and Teeth of Homo habilis* (Cambridge University Press, 1991).
- Wolpoff, M. H. *Human Evolution* (McGraw-Hill, New York, 1997).
- Clark, J. D. et al. African *Homo erectus*: old radiometric ages and young Oldowan assemblages in the Middle-Awash valley, Ethiopia. *Science* **264**, 1907–1909 (1994).

Acknowledgements. All fossils are stored at the Eritrea National Museum, Asmara; inventory numbers refer to the provisional catalogue. We thank M. Pickford and T. D. White for discussions about the fossil record; M. R. Gibling, I. P. Martini, P. Passerini and A. Turner for critical reading of an earlier manuscript; A. Kibrab (Eritrea Department of Mines) and Y. Libsekal (Eritrea National Museum) for assistance in Eritrea; and F. Heller (ETH Magnetic Laboratory, Zurich). This work was supported by the Italian CNR (Cultural Heritage and TRANSRIFT projects), the PeriTethys Programme, the European Commission, the University of Florence, and the Italian Ministry for Foreign Affairs.

Correspondence and requests for material should be addressed to E.A. (e-mail: abbate@cesit1.unifi.it).

Selective foraging behaviour of basking sharks on zooplankton in a small-scale front

David W. Sims & Victoria A. Quayle

Department of Biological Sciences and Plymouth Environmental Research Centre, University of Plymouth, Plymouth PL4 8AA, UK

The basking shark *Cetorhinus maximus* is the second largest fish species, attaining lengths of up to 11 m. During summer months in temperate coastal waters circumglobally, these sharks filter-feed on surface zooplankton^{1–4} near water-mass boundaries (fronts)^{5,6}; however, little else is known about their biology¹. Their foraging behaviour has not been investigated until now, although they have been described² as indiscriminate planktivores that are unlikely to orientate to specific plankton-rich waters. We have now tracked basking sharks responding to zooplankton gradients. We show that they are selective filter-feeders that choose the richest, most profitable plankton patches. They forage along thermal fronts and actively select areas that contain high densities of large zooplankton above a threshold density. They remain for up to 27 hours in rich patches that are transported by tidal currents and move between patches over periods of 1–2 days. We mapped feeding locations of these sharks in two years; the maps show that these sharks indicate broad shifts in front-located secondary production. Foraging behaviour of basking sharks therefore indicates the distribution, density and characteristics of zooplankton directly. This makes these sharks unique biological 'plankton recorders', with potential use as detectors of trends in abundance of zooplankton species that are influenced by climatic fluctuations of the North Atlantic Oscillation⁷.

Natural foraging behaviour of filter-feeding marine fishes is poorly understood because of the problems associated with tracking them and quantifying food abundance simultaneously. However, the basking shark spends long periods feeding at the waters' surface, a behaviour that allows study of this shark by tracking, observation of active feeding, and sampling of zooplankton prey. During May–July in 1996 and 1997, we tracked visually the movements of aggregations of feeding sharks ($n = 13$ groups) and the fine-scale foraging movements of individual basking sharks ($n = 14$), and measured zooplankton density on their swimming paths off Plymouth, southwest England (50° 16' N 004° 09' W). We also recorded the positions of basking sharks sighted in addition to those tracked. During 1996 and 1997, we observed 58 and 54 different sharks, respectively.

Foraging locations of basking sharks seen during May–July in 1996–97 were orientated along a seasonal boundary separating coastal inshore water off Plymouth from fully stratified water of the

northern side of the western English Channel (Fig. 1). The Plymouth front has been characterized by previous studies⁸⁻¹⁴ and lies approximately 8 km offshore, defined by the 14–16 °C contours that outcrop at the sea's surface^{14,15}. The front's approximate position in 1996 was determined from an infrared image of sea surface temperature (SST) taken on 6 June 1996, when we tracked 11 sharks.

A temperature transect across this position confirmed the presence of the front, with a gradient of 1.5 °C (between mean temperatures 12.3 °C and 13.8 °C) over a horizontal distance of 4.1 km, corroborating the remotely sensed determinations (Fig. 1). Close to, we could see the front by the presence of narrow strips of flat water (slicks) next to rippled water¹⁶⁻¹⁸.

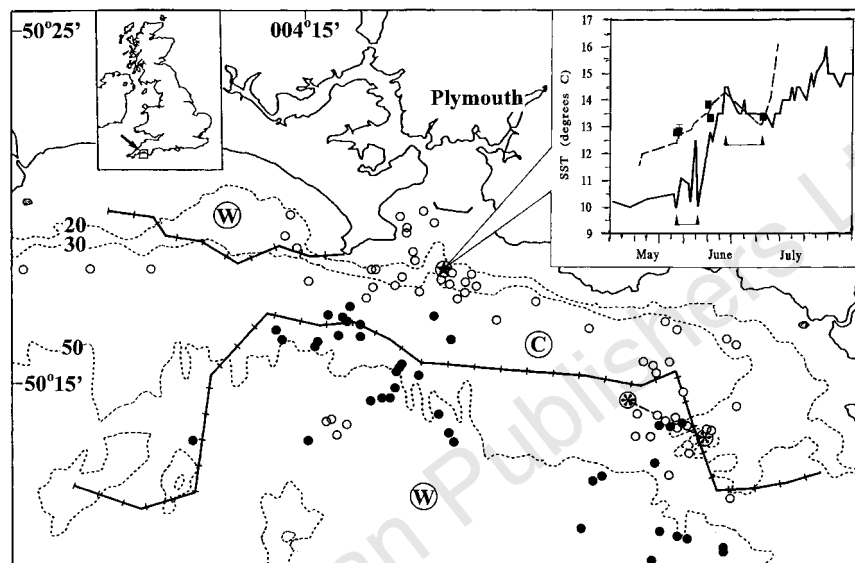


Figure 1 Distribution of surface-foraging basking sharks and changes in summer SST. The main figure shows the distribution of surface-foraging basking sharks in May–July, 1996 (open small circles, $n = 58$) and 1997 (filled circles, $n = 39$). Location points mark the extent of foraging locations. Bold ticked lines represent thermal-front boundaries between warm stratified water (W, 14–16 °C) and cooler mixed water (C, 12–13 °C) drawn from an infrared SST image from AVHRR channel 4 (NOAA-14 satellite) at 12:31, 6 June 1996. Tick marks are 1 km apart. Circled

asterisks joined by a dashed line indicate the SST (5 m depth) transect position across the front at 12:00–13:00, 17 June 1996. The top right inset shows the change in summer SST at the SST-sampling station, S1, in 1996 and 1997. Horizontal square brackets delimit periods of SST fluctuation. Filled squares represent mean SST while we were tracking individual basking sharks in 1997. Vertical bars denote ± 1 s.e.m. Dashed line, 1997 (2 m depth); complete line, 1996 (5 m depth). Dotted lines denote contour depths (20, 30, 50 m) and ticked lines are frontal boundaries.

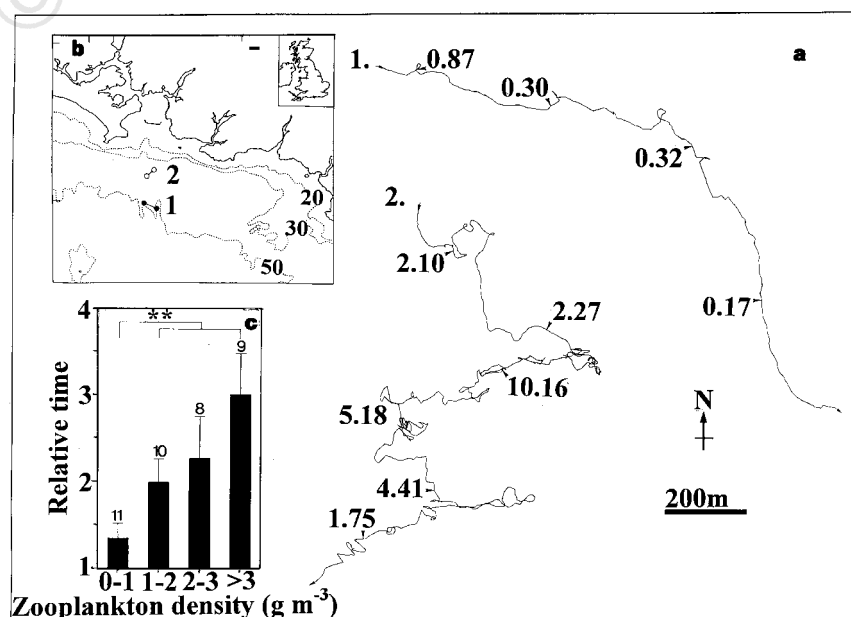


Figure 2 Feeding behaviour of basking sharks in relation to zooplankton density. **a**, Representative fine-scale-foraging tracks of two basking sharks responding to the zooplankton densities that they encountered. Track 1, swimming path of a shark (4 m L_T ; L_T is total length) tracked for 0.5 h (11:40–12:10, 30 May 1997). Track 2, foraging path for a different shark (4 m L_T) tracked for 1.7 h (08:20–10:02 h, 12 June 1997). Numbers along each track denote zooplankton densities sampled

in the sharks' paths in g wet weight per m^3 . **b**, Positions of trackings in relation to coastline and bathymetry off Plymouth. **c**, Relationship between zooplankton-density class and the time that 14 basking sharks spent filter-feeding within 25 m of the zooplankton-sample locations. A relative time of 1 equals 50 s. Vertical bars denote ± 1 s.e.m. Figures above bars represent n zooplankton samples. Two asterisks represent $P < 0.01$.

The distribution of basking sharks off Plymouth reflected broadly the front's known geographic position, indicating that the sharks may have concentrated their foraging activity along the boundary, generally in well mixed, cooler water (Fig. 1). However, in 1996, 91.5% of foraging locations were closest to the depth contours at 20–30 m, whereas in 1997 87.2% of locations were nearer to the 50-m contour, a further 4–7 km offshore (Fig. 1). Plymouth frontal sharpness varies according to the weather, as mixed inshore water becomes partly stratified, and hence warmer, in calm conditions^{13,14}. The predominance of strong easterly winds in May and early June 1997 may have extended the inshore zone of vertically mixed water, thereby locating the front further offshore. Later in the summer in both years, temperature fluctuations of 2.6–3.0 °C at the SST-sampling station, S1 (Fig. 1, inset), followed by large increases in SST, coincided with periods of calm weather, which would be consistent with decreased front sharpness as inshore water stratified. Sharks were seldom seen when high SSTs (indicative of front breakdown) were recorded at S1. Hence, foraging basking sharks may indicate weather-induced changes in location and persistence of the front between years, as well as indicating the broad position of the front.

Surface-feeding basking sharks followed convoluted swimming paths along tidal slicks associated with the front, exhibiting area-restricted searching (ARS) behaviour where zooplankton densities were measured to be high ($>1 \text{ g m}^{-3}$) (Fig. 2a, b). As a consequence of this localized feeding response, the 14 sharks tracked spent 48%,

69% and 123% longer time feeding in locations where zooplankton densities were 1–2, 2–3, and $>3 \text{ g m}^{-3}$, respectively, compared with the time spent in areas with $<1 \text{ g m}^{-3}$ zooplankton (Kruskal-Wallis test: $H_c = 11.22$, $\chi^2_{0.05,3} = 7.82$, $P < 0.025$; with multiple comparisons, $Q_{0.05,4} = 2.64$, $Q = 3.32$, $P < 0.01$; Fig. 2c). Densities of $<1 \text{ g m}^{-3}$ did not produce well defined ARS behaviour in six different sharks (Fig. 2c), which generally continued swimming on a relatively straight, non-convoluted course (see, for example, track 1 in Fig. 2a). Thus, zooplankton densities close to 1 g m^{-3} could represent the threshold below which it becomes increasingly less profitable to filter-feed.

Although time spent foraging by ten different sharks in zooplankton densities above the observed threshold showed a graded increase with productivity, mean times were not significantly different (multiple comparison: $Q = 1.51$, $P > 0.50$; Fig. 2c) because of greater temporal variability in ARS behaviour. We speculate that sharks already feeding on dense patches of zooplankton showed greater variability in food-orientated activity because prey was not limiting. Sharks feeding in zooplankton densities below 1 g m^{-3} exhibited less variable foraging responses (that is, they swam in a straight course), a behaviour that could act to increase the horizontal distance covered and consequently their chances of encountering new, more productive patches.

Basking sharks surface-feed in areas in which the dominant calanoid copepod prey, *Calanus helgolandicus*, is 2.5 times as numerous ($\sim 1,500$ organisms per m^3) and 50% longer ($\sim 2 \text{ mm}$),

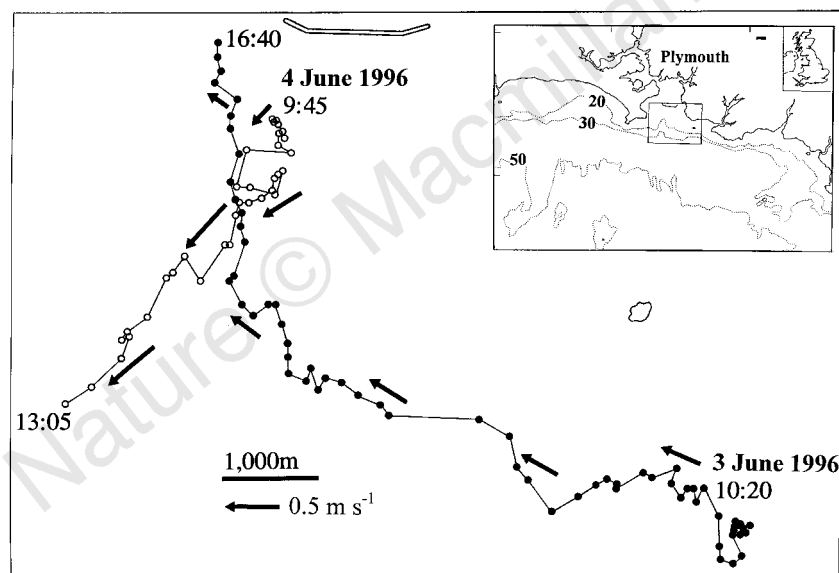


Figure 3 Movements of a feeding group of basking sharks over two consecutive days (3 June 1996, filled circles, and 4 June 1996, open circles). Position markers are given for every 5–10 min, with arrows denoting the direction of the tidal stream at hourly intervals during tracking. Tidal-stream speed is proportional to the length of the arrow.

Table 1 Group tracking data

Track date	Duration (h)	Number of sharks	Minimum distance covered (km)	Mean tidal-stream speed (m s^{-1})	Mean shark heading (degrees)	Mean tidal-stream direction (degrees)	Difference between shark heading and tidal-stream direction (degrees)
10.5.96	1.9	3	1.9	0.2	62.7	59.8	2.9
17.5.96	3.6	5	7.1	0.3	264.3	261.7	2.6
28.5.96	2.9	2	4.2	0.2	—	—	—
3.6.96	6.3	5	12.6	0.3	317.6	318.2	0.6
4.6.96	3.3	3	6.2	0.5	200.2	223.7	23.5
5.6.96	2.5	6	4.7	0.2	271.5	267.3	4.2
6.6.96	4.2	11	7.2	0.3	151.3	133.8	17.5
7.6.96	3.0	12	7.4	0.2	133.3	147.6	14.3
12.6.96	0.8	3	1.2	0.2	233.5	251.5	18.0
8.7.96	1.3	2	2.3	0.2	102.2	98.7	3.5
9.7.96	1.1	2	2.0	0.2	82.7	75.0	7.7
17.7.96	3.0	4	5.4	0.2	—	—	—
13.6.97	1.4	6	1.8	0.1	116.5	149.6	33.1

than in areas in which sharks do not feed¹⁹. In the feeding areas there are also fewer numbers of smaller zooplankton species, and therefore the biomass per cubic metre where sharks' forage is significantly increased¹⁹. These and our present results indicate that basking sharks may orientate to specific zooplankton prey by exhibiting ARS behaviour in response to fine-scale variations in zooplankton quantity and quality. In this way the basking sharks could locate the densest, most energy-rich patches. The sensory systems used by basking sharks to achieve these behaviours are unknown, but could involve electroreception²⁰ of copepod muscle activity and olfaction of dimethyl sulphide, which is produced by phytoplankton when grazed by zooplankton and is used as a foraging cue by some seabirds²¹.

Course headings of groups of foraging basking sharks followed closely the directions of tidal streams, indicating that they remained in, and tracked the movements of, productive patches transported by currents for at least 0.8–6.3 h (Fig. 3 and Table 1). Over distances travelled daily of 1.2–12.6 km, the headings of basking sharks compared with tidal-stream directions showed a mean difference of only 11.6° (± 3.2 s.e.m., $n = 11$; Table 1). One feeding group followed a productive patch as it was transported towards Plymouth Sound (Fig. 3), and, after one tidal cycle, was found to be tracking the patch back out to sea with the second outgoing tide. The minimum estimated time over which this feeding group remained in this patch was 27 h and the minimum distance that the patch and sharks moved was 18.8 km (track dates 3.6.96 and 4.6.96 in Table 1).

Our tracking studies also indicated movements of sharks between patches over periods of 1–2 days. Along the Plymouth front during 1996–97 (Fig. 1a), we relocated (separately) three sharks feeding in different patches 18–28 h after our first trackings and 5–11 km distant from initial foraging areas of the previous day. Two sharks that originally fed in the same patch moved in similar directions along a zooplankton gradient from low to higher density (range: 0.47–1.11 g m⁻³ to 1.06–1.43 g m⁻³), covering minimum distances of 9.5 km and 10.6 km in 27.6 h and 23 h respectively. The third shark was relocated feeding in slightly lower densities than 17.9 h earlier, and 4.7 km distant from the calculated new position of the first patch. These results indicate that basking sharks may move between patches, probably in response to low zooplankton densities encountered previously, minimizing travelling time between patches by using frontal boundaries to find successive patches in close proximity.

The areas where basking sharks foraged were also used by commercially important shoaling fishes. When tracking sharks along tidal slicks, large fish shoals appeared frequently on our echosounder, and fish were often seen thrashing the water's surface when feeding, sometimes doing so immediately ahead of feeding sharks. We identified large shoals of mackerel (*Scomber scombrus*), small whiting (*Merlangius merlangus*) and grey mullet (*Chelon labrosus*) near sharks. We also noted the activity of the northern gannet (*Sula bassana*) when tracking sharks. Around Britain and Ireland, gannets feed on shoaling fishes (length range 2.5–30.5 cm), primarily herring (*Clupea harengus*), sprat (*Sprattus sprattus*), mackerel, cod (*Gadus morhua*), haddock (*Melanogrammus aeglefinus*) and small whiting²². We saw gannets diving and resting at the surface within the sharks' foraging areas during 22 of our 27 trackings (82%). This indicates that western Channel fronts are important hydrographic features for commercial teleosts as well as sharks.

Thus, basking sharks foraged for longer periods in patches with high zooplankton density above a threshold density of 1 g m⁻³; these patches contained more numerous large prey. The highest densities produced the most prolonged ARS responses. Basking sharks also tracked tidally transported planktonic patches and moved between patches within fronts over broader temporal and spatial scales (10–20 km in 1–2 days). Mapping their foraging locations indicated the position of a seasonal front, and indicated a westward shift in

biological front location between 1996 and 1997, perhaps because of weather conditions. Basking sharks exhibit predictable ARS behaviour in response to specific zooplankton densities and characteristics¹⁹ and follow and move between tidally transported patches in fronts, identifying the basking shark as a selective forager on zooplankton assemblages; this behaviour was previously unknown.

As fronts show enhanced production of plankton^{16,23,24}, attended by increased fish stocks^{15,17,25}, identifying frontal regions from new satellite-derived ocean-colour data of phytoplankton abundance (for example, using SeaWiFS) may help indirectly in studies of zooplankton and fish distribution²⁶. However, basking sharks integrate a planktivorous fish's behaviour with zooplankton abundance directly. In addition, fluctuations in the abundance of two *Calanus* species (upon which basking sharks feed) are closely linked to the state of the North Atlantic Oscillation⁷, emphasizing the influence of climate variations on zooplankton. Thus we propose that basking sharks are potentially useful detectors of the distribution, density and characteristics of these zooplankton in fronts, and could provide cheap, high-trophic-level biological indication of fluxes in zooplankton assemblages that are affected by oceanographic and climatic fluctuations in the North Atlantic. □

Methods

Fish tracking. We located surface-foraging basking sharks by visual survey from a 10-m research vessel. We stayed within 10–50 m of sharks during tracking, except when zooplankton sampling, plotting our location using a global positioning system (GPS) (Garmin 120S) every 5–10 min. Fine-scale-foraging routes were determined by vectors, using the time each shark spent on successive course headings (obtained from a hand-held digital compass, with headings recorded in real time on audiotape) multiplied by the mean feeding-swimming speed, to calculate estimated distance travelled between course deviations. Feeding-cruising speeds were measured for six sharks using onboard differential GPS (Valsat 03, MLR Electronics, France) and a digital flowmeter (Solomat 520c) attached to a towed propeller (General Oceanics 2030R, USA). Under the feeding conditions observed, basking shark swimming speeds remained relatively constant at 0.97 m s⁻¹ (± 0.03 s.e.m., $n = 59$ determinations; mean shark length = 4.7 m ± 1.1 s.e.m.). Tracks were reconstructed graphically using the compass headings and data on distance travelled. Data on tidal-stream direction and speed for each track were calculated from data given on Admiralty Chart number 1613 (Crown Copyright 1995, Hydrographic Office, UK) made at tidal-stream recording stations A–D, located at: A, 50° 18.3 N 4° 10.8 W; B, 50° 18.3 N 4° 07.7 W; C, 50° 12.5 N 4° 05.2 W; and D, 50° 07.7 N 3° 55.2 W. Recording stations A and B are positioned on the 20-m contour in the western and eastern approaches to Plymouth Sound, respectively. The SST sampling station (S1) lies at a position equidistant to A and B at a distance of 1.8 km. Tidal-stream recording stations C and D are located near the frontal boundary.

Zooplankton and water-temperature sampling. Zooplankton samples were taken within 3 m of feeding paths of sharks and analysed using described methods^{1,19}. One to six samples were taken per individual fish track. Water temperature was taken with a Yellow Springs Instruments model 58 meter and a standard mercury thermometer with minimum–maximum markers.

Track analysis. The time sharks spent foraging near measured zooplankton densities was quantified by drawing a circular area of interest (AOI) around each sample location on each shark's graphically reconstructed foraging track. The AOI radius was set at 25 m; this area was to analyse activity at each zooplankton-sample location on all tracks. The length of track falling within each AOI, indicating the degree of localization of foraging responses, was measured (in mm) on track reconstructions and converted to time in seconds, as distance covered at a uniform cruising speed was directly proportional to time. The degree of dispersion of foraging time between zooplankton-density classes was examined with a Kruskal-Wallis test with tied ranks, performed with non-parametric multiple comparisons for unequal sample sizes²⁷.

Received 9 October 1997; accepted 9 March 1998.

1. Sims, D. W., Fox, A. M. & Merrett, D. A. Basking shark occurrence off south-west England in relation to zooplankton abundance. *J. Fish Biol.* **51**, 436–440 (1997).

- Matthews, L. H. & Parker, H. W. Notes on the anatomy and biology of the basking shark (*Cetorhinus maximus* (Gunner)). *Proc. Zool. Soc. Lond.* **120**, 555–576 (1950).
- Kenney, R. D., Owen, R. E. & Winn, H. E. Shark distributions off the northeast United States from marine mammal surveys. *Copeia* **1985**, 220–223 (1985).
- Berrow, S. D. & Heardman, C. The basking shark *Cetorhinus maximus* (Gunnerus) in Irish waters—patterns of distribution and abundance. *Proc. R. Irish Acad. B* **94**, 101–107 (1994).
- Priede, I. G. A basking shark (*Cetorhinus maximus*) tracked by satellite together with simultaneous remote sensing. *Fish. Res.* **2**, 201–216 (1984).
- Choy, B. K. & Adams, D. H. An observation of a basking shark, *Cetorhinus maximus*, feeding along a thermal front off the east-central coast of Florida. *Florida Sci.* **58**, 313–319 (1995).
- Fromentin, J.-M. & Planque, B. *Calanus* and environment in the eastern North Atlantic. II. Influence of the North Atlantic Oscillation on *C. finmarchicus* and *C. helgolandicus*. *Mar. Ecol. Prog. Ser.* **134**, 111–118 (1996).
- Harvey, H. W., Cooper, L. H. N., Lebour, M. V. & Russell, F. S. Plankton production and its control. *J. Mar. Biol. Assoc. UK* **20**, 407–441 (1935).
- Southward, A. J. On changes of sea temperature in the English Channel. *J. Mar. Biol. Assoc. UK* **39**, 449–458 (1960).
- Armstrong, F. A. J., Butler, E. I. & Boalch, G. T. Hydrographic and nutrient chemistry surveys in the western English Channel during 1963 and 1964. *J. Mar. Biol. Assoc. UK* **52**, 915–930 (1972).
- Southward, A. J. & Butler, E. I. A note on further changes of sea temperature in the Plymouth area. *J. Mar. Biol. Assoc. UK* **52**, 931–937 (1972).
- Southward, A. J. & Demir, N. The abundance and distribution of eggs and larvae of some teleost fishes off Plymouth in 1969 and 1970. *J. Mar. Biol. Assoc. UK* **52**, 987–996 (1972).
- Southward, A. J. The western English Channel—an inconstant ecosystem? *Nature* **285**, 361–366 (1980).
- Southward, A. J. Fluctuations in the 'indicator' chaetognaths *Sagitta elegans* and *Sagitta setosa* in the western Channel. *Oceanol. Acta* **7**, 229–239 (1984).
- Robinson, I. S. *Satellite Oceanography* (Wiley, Chichester, 1994).
- Wolanski, E. & Hamner, W. M. Topographically controlled fronts in the ocean and their biological significance. *Science* **241**, 177–181 (1988).
- Kingsford, M. J., Wolanski, E. & Choat, J. H. Influence of tidally induced fronts and Langmuir circulations on distribution and movements of presettlement fishes around a coral reef. *Mar. Biol.* **109**, 167–180 (1991).
- Le Fèvre, J. Aspects of the biology of frontal systems. *Adv. Mar. Biol.* **23**, 163–299 (1986).
- Sims, D. W. & Merrett, D. A. Determination of zooplankton characteristics in the presence of surface feeding basking sharks, *Cetorhinus maximus*. *Mar. Ecol. Prog. Ser.* **158**, 297–302 (1997).
- Kalmijn, A. J. Electric and magnetic field detection in elasmobranch fishes. *Science* **218**, 916–918 (1982).
- Nevitt, G. A., Veit, R. R. & Kareiva, P. Dimethyl sulphide as a foraging cue for Antarctic Procellariiform seabirds. *Nature* **376**, 680–682 (1995).
- Cramp, S. et al. *Handbook of the Birds of Europe the Middle East and North Africa: The Birds of the Western Palearctic Vol. 1* (Oxford Univ. Press, Oxford, 1977).
- Pingree, R. D., Pugh, P. R., Holligan, P. M. & Forster, G. R. Summer phytoplankton blooms and red tides along tidal fronts in the approaches to the English Channel. *Nature* **258**, 672–677 (1975).
- Yoder, J. A. et al. A line in the sea. *Nature* **371**, 689–692 (1994).
- Fiedler, P. C., Smith, G. B. & Laurs, R. M. Fisheries applications of satellite data in the eastern north Pacific. *Mar. Fish. Rev.* **46**, 1–13 (1984).
- Southwell, K. Oceanography: Shades of the sea. *Nature* **389**, 444 (1997).
- Zar, J. H. *Biostatistical Analysis* (Prentice-Hall, Englewood Cliffs, 1984).

Acknowledgements. The research programme was supported by the Nature Conservancy Council for England (English Nature) and the Plymouth Environmental Research Centre. We thank A. Fox, B. Broughton, D. Murphy, A. Giles, D. Üren, R. Harris, P. Ede and R. Hopgood for their assistance in research at sea; D. Merrett for help in zooplankton analysis; and the late F. Carey and D. Nelson, whose work helped to inspire the principal investigator (D.W.S.) to begin these studies. Satellite images of SST were supplied by the NERC Satellite Receiving Station.

Correspondence and requests for materials should be addressed to D.W.S. (e-mail: dwsims@plymouth.ac.uk).

Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes

Lisa R. Moore^{†‡}, Gabrielle Rocap^{†‡§} & Sallie W. Chisholm^{†‡}

[†]Department of Civil and Environmental Engineering, [‡]Department of Biology, 48–425 Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

[§]MIT/Woods Hole Oceanographic Institute Joint Program in Oceanography, Woods Hole, Massachusetts 02543, USA

* These authors contributed equally to this work

The cyanobacterium *Prochlorococcus*^{1,2} is the dominant oxygenic phototroph in the tropical and subtropical regions of the world's oceans^{1,3,4}. It can grow at a range of depths over which light intensities can vary by up to 4 orders of magnitude. This broad depth distribution has been hypothesized to stem from the coexistence of genetically different populations adapted for growth at high- and low-light intensities^{4–6}. Here we report

direct evidence supporting this hypothesis, which has been generated by isolating and analysing distinct co-occurring populations of *Prochlorococcus* at two locations in the North Atlantic. Co-isolates from the same water sample have very different light-dependent physiologies, one growing maximally at light intensities at which the other is completely photoinhibited. Despite this ecotypic differentiation, the co-isolates have 97% similarity in their 16S ribosomal RNA sequences, demonstrating that molecular microdiversity, commonly observed in microbial systems^{7–12}, can be due to the coexistence of closely related, physiologically distinct populations. The coexistence and distribution of multiple ecotypes permits the survival of the population as a whole over a broader range of environmental conditions than would be possible for a homogeneous population.

Using sea-going flow cytometry for studying picoplankton populations, we and others^{4,13,14} have observed multiple populations of *Prochlorococcus* in single water samples, as distinguished by their chlorophyll fluorescence intensities. These populations could be derived from the mixing together of genetically identical *Prochlorococcus* cells which have acclimated to different past light

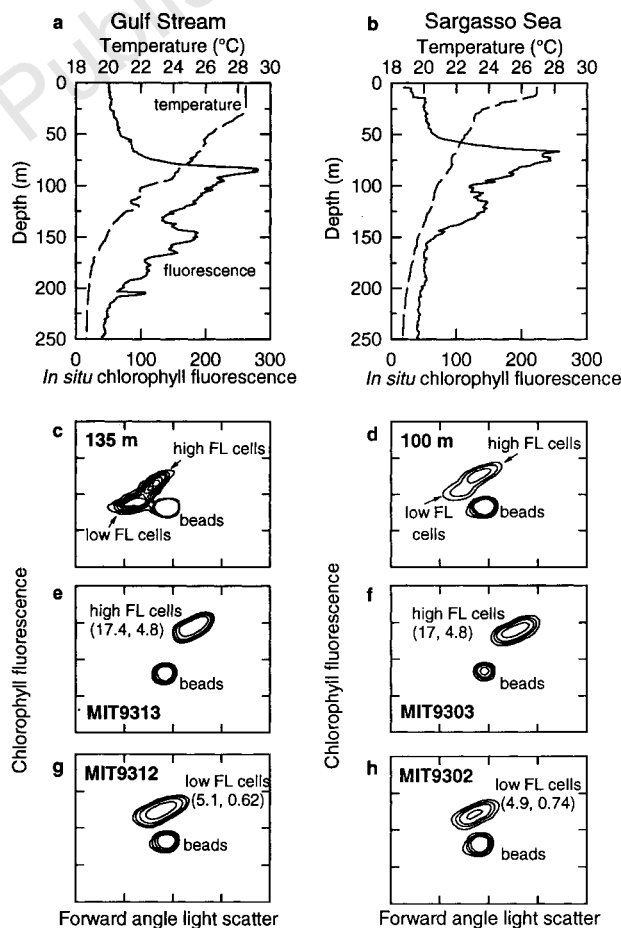


Figure 1 Properties of the euphotic zone and flow cytometric signatures of the *Prochlorococcus* populations and isolates. **a, b**, The physical features of the water columns were similar at the Gulf Stream station, 37° 30.8'N, 68° 14.4'W, and the Sargasso Sea station, 34° 45.5'N, 66° 11.1'W. **c, d**, Flow cytometry signatures of coexisting *Prochlorococcus* populations from 135 m in the Gulf Stream and 100 m in the Sargasso Sea from which the isolates were obtained. **e–h**, Flow cytometry signatures of the cultured isolates maintained at an irradiance of 9 $\mu\text{mol l}^{-1} \text{s}^{-1}$. Numbers in parentheses refer to the mean chlorophyll fluorescence per cell (FL) and FALS per cell. Differences in the absolute values of the flow cytometry parameters between natural populations and isolates result from unmatched growth conditions.

and/or nutrient regimes, or they could represent genetically distinct populations, adapted for optimal growth under different conditions but co-occurring in environments where conditions are favourable for both. The latter interpretation has been suggested by indirect evidence from field studies^{4,13,14}, from physiological diversity among *Prochlorococcus* isolates⁶, and from studies of the molecular phylogenetic diversity of marine cyanobacteria, which have revealed a number of closely related gene sequences⁷⁻⁹. This high microdiversity has also been observed for marine bacterioplankton^{7,8,10} and microorganisms in other environments^{11,12}, although the ecological significance has remained unclear. Differences in RNA polymerase gene sequences have been correlated with pigment content in marine *Synechococcus*¹⁵; however, molecular microdiversity observed in the more widely used 16S rRNA gene has not been connected directly to physiological function in any group of microorganisms from a particular habitat. The multiple *Prochlorococcus* populations observed by flow cytometry provided an excellent opportunity for studying these relationships.

In two locations of the North Atlantic Ocean at the junction of two subsurface chlorophyll fluorescence maximum layers, we used flow cytometry to sort the high- and low-fluorescence cells from coexisting *Prochlorococcus* populations and then brought them into culture (Fig. 1). After several years of growing under identical conditions, the cells sorted from a Gulf Stream water sample (isolates MIT9312 and MIT9313) and a Sargasso Sea water sample (isolates MIT9302 and MIT9303) have maintained their differences in chlorophyll fluorescence and forward-angle light scatter (FALS) per cell (Fig. 1e-h), indicating that the phenotypic differences originally observed in these coexisting populations were due to genetic differences.

We hypothesized that the high- and low-fluorescence *Prochlorococcus* are low- and high-light adapted, respectively, because differences in flow cytometry parameters are correlated with differences in photophysiology between *Prochlorococcus* SS120,

isolated from the Sargasso Sea, and MED4, isolated from the Mediterranean Sea⁶. To test this hypothesis, we compared these four isolates with respect to their light-dependent growth response. The pattern of growth rate as a function of irradiance differs greatly between each coexisting pair (Fig. 2a, b). The most striking difference is that the high-fluorescence isolates, MIT9303 and MIT9313, show complete photoinhibition of growth at irradiances where the low-fluorescence isolates, MIT9302 and MIT9312, are growing at or close to maximal rates (Fig. 2a, b). Additionally, MIT9303 and MIT9313 grow better at lower light than their co-isolates (Fig. 2a, b).

To understand better the physiological underpinnings of the differential growth response of the isolates, we compared their pigment content, absorption properties and photosynthetic performance when grown at low irradiance. The high-fluorescence isolates, MIT9303 and MIT9313, contain higher concentrations of divinyl chlorophylls *a* and *b* (chl *a*₂ and chl *b*₂) relative to their low-fluorescence co-isolates MIT9302 and MIT9312 (Table 1). They also contain higher ratios of chl *b*₂/*a*₂, resulting in higher spectrally weighted average chlorophyll *a*₂-specific absorption coefficients, $\bar{a}_{chl}$

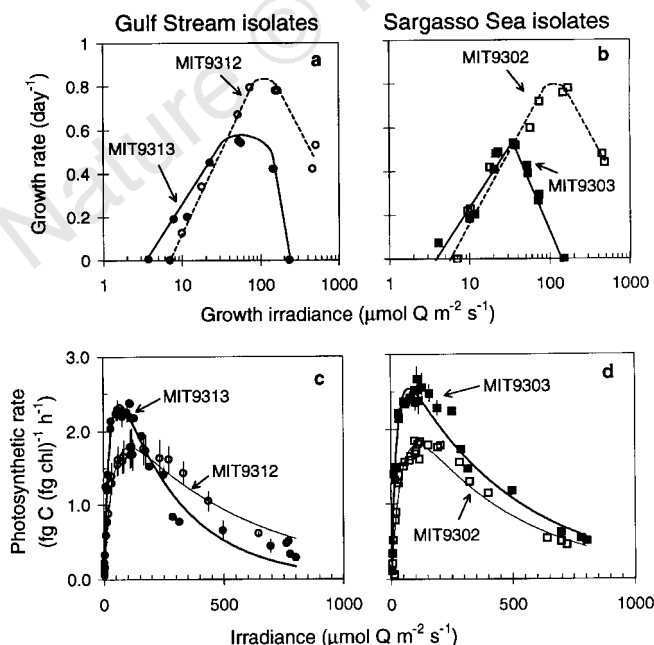


Figure 2 Growth and photosynthetic rate as a function of irradiance for the *Prochlorococcus* isolates shown in Fig. 1. **a, b**, Specific growth rate as a function of growth irradiance. **c, d**, Photosynthetic rate (normalized to chl *a*₂) as a function of irradiance for the four isolates when grown at an irradiance of 9 μmol Q m⁻² s⁻¹. Symbols and error bars correspond to the mean ± 1 s.e. of duplicate measurements.

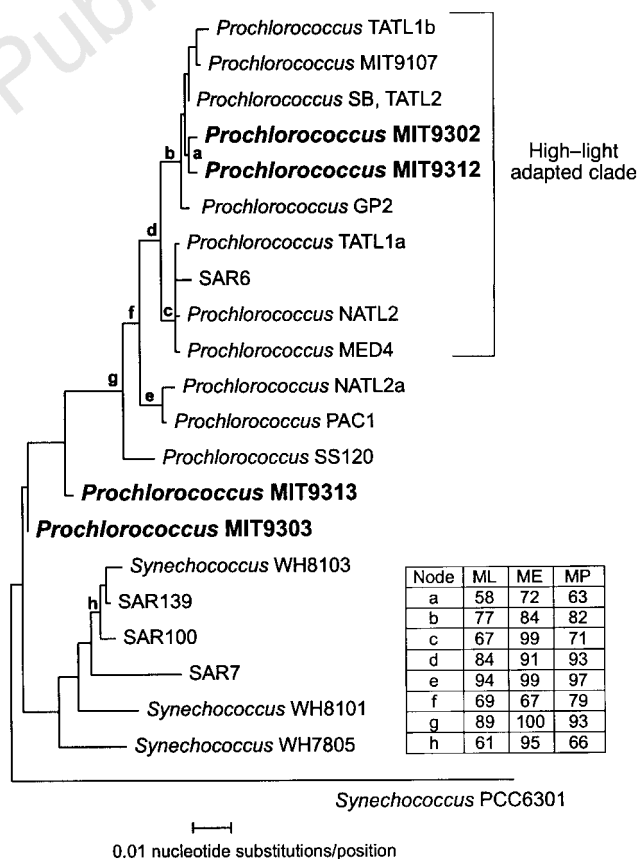


Figure 3 Phylogenetic relationships of *Prochlorococcus* and *Synechococcus* isolates and environmental sequences from the Sargasso Sea ('SAR') inferred from 16S rRNA sequences. One of the three most probable trees found in a maximum likelihood analysis is presented. (Only the relative branching orders of *Prochlorococcus* isolates MED4, TATL1a and NATL2 and the environmental sequence SAR6 differ between the three trees.) Bootstrap proportions are percentages of 100 resampled data sets from analyses of maximum likelihood (ML), distance (using minimum evolution as the objective function (ME)) and maximum parsimony (MP); values below 50 are not shown. The freshwater cyanobacterium *Synechococcus* PCC 6301 was used to root the tree. Analyses with additional members of the cyanobacterial radiation resulted in essentially similar branching orders (data not shown).

Table 1 Physiological characteristics of *Prochlorococcus* isolates grown at low growth irradiance of $9 \mu\text{mol Q m}^{-2} \text{s}^{-1}$

Parameter	Gulf Stream		Sargasso Sea	
	MIT9312	MIT9313	MIT9302	MIT9303
Pigments				
chl a_2	3.15	5.1	3.2	4.2
chl b_2	2.08	6.0	2.1	5.0
chl b_2/a_2	0.661	1.175	0.65	1.19
Absorption				
$\bar{a}_{chl}$	0.0152	0.0213	0.0147	0.0232
Photosynthetic				
α_{chl}	0.057	0.118	0.055	0.125
P_{max}	1.8	2.4	1.8	2.54
ϕ_{max}	0.086	0.128	0.086	0.125

Chl a_2 and chl b_2 are in units of fg per cell; chl b_2/a_2 are on a weight/weight basis. $\bar{a}_{chl}$ is presented in units of $(\text{m}^2 (\text{mg chl } a_2)^{-1})$. Photosynthetic parameters are in the following units: $\alpha_{chl} = \text{fg C} (\text{fg chl } a_2)^{-1} \text{h}^{-1}$ ($\mu\text{mol Q m}^{-2} \text{s}^{-1}$); $P_{max} = \text{fg C} (\text{fg chl } a_2)^{-1} \text{h}^{-1}$; and $\phi_{max} = \text{mol carbon} (\text{mol Q})^{-1}$. All values are means of duplicate cultures with coefficients of variance $\leq 10\%$.

(Table 1). Furthermore, the higher chl b_2/a_2 ratio of MIT9303 and MIT9313 drives a 2.1–2.3 times higher chl a_2 -specific photosynthetic efficiency (α_{chl}) and results in higher maximum quantum yields (ϕ_{max}) relative to MIT9302 and MIT9312 (Fig. 2c, d; Table 1). Thus, we conclude that the isolates MIT9303 and MIT9313 are low-light adapted and MIT9302 and MIT9312 are high-light adapted.

We next explored whether the physiology of these four isolates is related to their phylogeny by sequencing their 16S rRNA genes. Although they were isolated from different waters, the two high-light adapted isolates, MIT9302 and MIT9312, have very high sequence similarity (99.7%), as do the two low-light adapted isolates, MIT9303 and MIT9313 (99.2%). The two co-isolates from each region are more disparate than the high- and low-light adapted pairs: the Sargasso Sea co-isolates MIT9302 and MIT9303 are only 97.3% similar, whereas the Gulf Stream co-isolates MIT9312 and MIT9313 are 97.7% similar. The divergence between the co-isolates is meaningful because their degrees of similarity are comparable to those between the low-light adapted *Prochlorococcus* isolates and marine isolates of phycobilisome-containing *Synechococcus* (96.9–98%). A phylogenetic tree constructed from the coexisting *Prochlorococcus* and other cyanobacterial sequences shows that the high-light adapted MIT9312 and MIT9302 cluster with the high-light adapted MED4 and two isolates from the Pacific, SB and GP2, which have low ratios of chl b/a_2 (ref. 16) (Fig. 3). The monophyly of this 'high-light adapted clade'¹⁷ is well supported in all analyses. The low-light adapted MIT9303 and MIT9313 are on separate, more basal branches of the tree (Fig. 3). Thus, the phylogenetic relationship is correlated with the phenotypic relationship.

On the basis of the physiology and molecular phylogeny of these co-isolates, the previously characterized isolates, SS120 and MED4^{6,17–19}, and four additional isolates from the Pacific Ocean (ref. 20 and G.R. and S.W.C., unpublished results), it appears that at least two distinct ecotypes can be identified for isolates of the genus *Prochlorococcus*. This is analogous to two types of marine *Synechococcus*, which have different ratios of phycourobilin to phycoerythrobilin²¹, and which appear to be distributed differently on the basis of their relative light harvesting abilities²². It has been suggested that low-light adapted, high chl b_2/a_2 *Prochlorococcus* predominate in the deeper portion of the euphotic zone where nutrients are abundant, and that high-light adapted, low chl b_2/a_2 *Prochlorococcus* predominate in the surface where nutrients are typically limiting^{4–6}. This distribution of multiple *Prochlorococcus* ecotypes in the same water column would result in greater integrated production than could be achieved by a single ecotype. For example, if we estimate primary production ($P = \bar{a}_{chl} \phi_{max} C_c E$) for the same number of cells (C_c) in the low light of the deep euphotic zone ($E = 10 \mu\text{mol Q m}^{-2} \text{s}^{-1}$) (see Methods), using the strain-

specific values of $\bar{a}_{chl}$, ϕ_{max} and chl a_2 per cell (C_c), then the contribution of the low-light adapted MIT9313 to primary production would be 3.4 times higher than that of its high-light adapted coisolate MIT9312. Thus, the use of a single set of physiological parameters will result in errors in the estimates of *Prochlorococcus* primary production in the oceans.

Our results show that high phylogenetic microdiversity observed for coexisting marine cyanobacterial picoplankton^{7–9} is due, in part, to physiological diversity within the *Prochlorococcus* population. Although studies using allozyme banding patterns have demonstrated ecotypic differentiation in eukaryotic phytoplankton^{23,24}, we have gone a step further in that we have specifically linked physiological diversity to small differences in 16S rRNA sequences. Thus, high microdiversity at the 16S rRNA locus observed in other microbial communities examined primarily by comparative sequence analysis may also reflect physiologically distinct microbial populations. The existence of multiple ecotypes in a particular environment may be a general phenomenon in the microbial world, allowing for survival over a broader range of conditions than could be achieved by a physiologically and genetically homogeneous population. □

Methods

Flow cytometry and culturing. Field populations were identified as *Prochlorococcus* on the basis of the characteristic low chlorophyll fluorescence and FALS signature on the flow cytometer¹. Populations were designated as 'double' when bimodal signatures were visible, defined by different fluorescence and FALS signals (Fig. 1). Coexisting *Prochlorococcus* populations were sorted from each other into sterile test tubes of modified K/10-Cu media with 1.17 μM EDTA⁶ using an EPICS 753 flow cytometer (Coulter Corp.). Cells were grown at 21 °C and 17 $\mu\text{mol quanta m}^{-2} \text{s}^{-1}$ blue light (14:10 light:dark) immediately after isolation on the ship, and back in the laboratory at 24 °C and varying light intensities obtained using neutral density filters and cool-white fluorescent bulbs. For photosynthesis, pigment and absorption measurements, cells were grown at a low irradiance ($9 \mu\text{mol Q m}^{-2} \text{s}^{-1}$), where differences between SS120 (CCMP1375) and MED4 (CCMP1378) were apparent⁶. Cell counts for growth rate calculations and normalization of pigment data were obtained using a FACScan flow cytometer (Becton-Dickinson), and mean chlorophyll fluorescence per cell and FALS are presented relative to 0.474 μm yellow-green fluorescent beads (Polysciences). Flow cytometric data were analysed using 'CytoPC' software provided by D. Vaultot (Station Biologique, Roscoff, France).

Photosynthesis, pigment and absorption measurements. Photosynthesis-irradiance measurements were carried out on exponentially growing cells spiked with $\text{NaH}^{14}\text{CO}_3$ (0.1 $\mu\text{Ci ml}^{-1}$ culture; specific activity between 100,000 and 200,000 d.p.m.). The spiked culture was dispensed into glass scintillation vials (1 ml each) and incubated at 24 °C for 45 min over a range of irradiances obtained using very high output/daylight spectrum fluorescence bulbs attenuated with neutral density filters and measured with a 4 π quantum light meter (model QSL-100, Biospherical Instruments, San Diego, CA). Photosynthesis was terminated and unincorporated ^{14}C removed by adding 2 M HCl (100 μl) and shaking for ~2 h under a fume hood. Samples were then analysed by scintillation spectrometry. The photosynthetic parameters, P_{max} and α , were obtained by fitting the results to the equation from ref. 25, using the curve-fitting program in SigmaPlot (Jandel Scientific), and normalized to chl a_2 and chl b_2 which were measured spectrophotometrically²⁶. *In vivo* chlorophyll a_2 -specific absorption, $\bar{a}_{chl}(\lambda)$ ($\text{m}^2 (\text{mg chl } a_2)^{-1}$), was obtained using an opal diffuser on a Beckman DU-640 spectrophotometer, and the weighted average ($\bar{a}_{chl}^*$) was calculated over the photosynthetically available radiation range (400–700 nm). Maximum quantum yield, ϕ_{max} , is calculated as the ratio of α_{chl} to $\bar{a}_{chl}$.

Molecular phylogenetic analysis. Genomic DNA was isolated from cultures according to standard methods²⁷. 16S rDNA was amplified using the general eubacterial primers 8-27f (AGAGTTTGATCCTGGCTCAG) and 1504-1486r (CTTGTTACGACTTCACCCC). Polymerase chain reactions (PCRs) were performed in quintuplicate, using the high-fidelity polymerase *Pfu* (Stratagene). Reactions were pooled and purified using QIA quick kit (QIAGEN,

Chatsworth, CA) and cloned using the pCR-Script kit (Stratagene). A minimum of 8 clones from each culture were sequenced on an automated sequencer (LI-COR, Lincoln, NE). Sequences were aligned manually with other marine cyanobacterial sequences available in the Ribosomal Database Project²⁸ using the Genetic Data Environment²⁹. The sequence for MIT9303 obtained previously from a PCR product¹⁷ is contained within the sequence we report here. A total of 1,094 unambiguously aligned and determined nucleotides were used in the analyses. Phylogenetic analyses used PAUP* (version 4.0d47, provided by D. Swofford). For both distance and maximum likelihood analyses the model of nucleotide substitution used was the Hasegawa Kishino Yana 1985 model. Nucleotide frequencies and the transition transversion ratio were estimated from the data.

Received 11 November 1997; accepted 26 March 1998.

- Chisholm, S. W. *et al.* A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* **334**, 340–343 (1988).
- Chisholm, S. W. *et al.* *Prochlorococcus marinus* nov. gen. nov. sp.: an oxyphototrophic marine prokaryote containing divinyl chlorophyll *a* and *b*. *Arch. Microbiol.* **157**, 297–300 (1992).
- Olson, R. J., Chisholm, S. W., Zettler, E. R., Altabet, M. A. & Dusenberry, J. A. Spatial and temporal distributions of prochlorophyte picoplankton in the North Atlantic Ocean. *Deep Sea Res.* **37**, 1033–1051 (1990).
- Campbell, L. & Vaulot, D. Photosynthetic picoplankton community structure in the subtropical North Pacific Ocean near Hawaii (station ALOHA). *Deep Sea Res.* **40**, 2043–2060 (1993).
- Goericke, R. & Repeta, D. J. Chlorophylls *a* and *b* and divinyl chlorophylls *a* and *b* in the open subtropical North Atlantic Ocean. *Mar. Ecol. Prog. Ser.* **101**, 307–313 (1993).
- Moore, L. R., Goericke, R. & Chisholm, S. W. Comparative physiology of *Synechococcus* and *Prochlorococcus*: influence of light and temperature on growth, pigments, fluorescence and absorptive properties. *Mar. Ecol. Prog. Ser.* **116**, 259–275 (1995).
- Giovannoni, S. J., Britschgi, T. B., Moyer, C. L. & Field, K. G. Genetic diversity in Sargasso Sea bacterioplankton. *Nature* **345**, 60–63 (1990).
- Fuhrman, J. A., McCallum, K. & Davis, A. A. Phylogenetic diversity of subsurface marine microbial communities from the Atlantic and Pacific Oceans. *Appl. Environ. Microbiol.* **59**, 1294–1302 (1993).
- Palenik, B. Cyanobacterial community structure as seen from RNA polymerase gene sequence analysis. *Appl. Environ. Microbiol.* **60**, 3212–3219 (1994).
- DeLong, E. F., Wu, K. Y., Prézélin, B. B. & Jovine, R. V. M. High abundance of Archaea in Antarctic marine picoplankton. *Nature* **371**, 695–697 (1994).
- Ward, D., Weller, M. R. & Bateson, M. M. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. *Nature* **345**, 63–65 (1990).
- Amann, R., Snaird, J., Wagner, M., Ludwig, W. & Schleifer, K.-H. In situ visualization of high genetic diversity in a natural microbial community. *J. Bacteriol.* **178**, 3496–3500 (1996).
- Veldhuis, M. J. W. & Kraay, G. W. Cell abundance and fluorescence of picoplankton in relation to growth irradiance and nitrogen availability in the Red Sea. *Neth. J. Sea Res.* **31**, 135–145 (1993).
- Partensky, F., Blanchot, J., Lantoin, F., Neveaux, J. & Marie, D. Vertical structure of picophytoplankton at different trophic sites of the subtropical northeastern Atlantic Ocean. *Deep Sea Res.* **43**, 1191–1213 (1996).
- Toledo, G. & Palenik, B. *Synechococcus* diversity in the California Current as seen by RNA polymerase (*rpoC1*) gene sequences of isolated strains. *Appl. Environ. Microbiol.* **63**, 4298–4303 (1997).
- Shimada, A., Maruyama, T. & Miyachi, S. Vertical distributions and photosynthetic action spectra of two oceanic picophytoplankters, *Prochlorococcus marinus* and *Synechococcus* sp. *Mar. Biol.* **127**, 15–23 (1996).
- Urbach, E., Scanlan, D. J., Distel, D. L., Waterbury, J. B. & Chisholm, S. W. Rapid diversification of marine picophytoplankton with dissimilar light-harvesting structures inferred from sequences of *Prochlorococcus* and *Synechococcus* (Cyanobacteria). *J. Mol. Evol.* **46**, 188–201 (1998).
- Partensky, F., Hoepffner, N., Li, W. K. W., Ulloa, O. & Vaulot, D. Photoacclimation of *Prochlorococcus* sp. (Prochlorophyta) strains isolated from the north Atlantic and the Mediterranean Sea. *Plant Physiol.* **101**, 285–296 (1993).
- Scanlan, D. J., Hess, W. R., Partensky, F., Newman, J. & Vaulot, D. High degree of genetic variation in *Prochlorococcus* (Prochlorophyta) revealed by RFLP analysis. *Eur. J. Phycol.* **31**, 1–9 (1996).
- Moore, L. R. & Chisholm, S. W. Photophysiology of the marine cyanobacterium *Prochlorococcus*: ecotypic differences among cultured isolates. *Limnol. Oceanogr.* (submitted).
- Waterbury, J. B., Watson, S. W., Valois, F. W. & Franks, D. G. Biological and ecological characterization of the marine unicellular cyanobacterium *Synechococcus*. *Can. Bull. Fish. Aquat. Sci.* **214**, 71–120 (1986).
- Olson, R. J., Chisholm, S. W., Zettler, E. R. & Armbrust, E. V. Analysis of *Synechococcus* pigment types in the sea using single and dual beam flow cytometry. *Deep Sea Res.* **35**, 425–440 (1988).
- Gallagher, J. C. Population genetics of *Skeletonema costatum* (Bacillariophyceae) in Narragansett Bay. *J. Phycol.* **16**, 464–474 (1980).
- Gallagher, J. C. Physiological variation and electrophoretic banding patterns of genetically different seasonal populations of *Skeletonema costatum* (Bacillariophyceae). *J. Phycol.* **18**, 148–162 (1982).
- Platt, T., Gallegos, C. L. & Harrison, W. G. Photoinhibition of photosynthesis in natural assemblages of marine phytoplankton. *J. Mar. Res.* **38**, 687–701 (1980).
- Jeffrey, S. W. & Humphrey, G. F. New spectrophotometric equations for determining chlorophyll *a*, *b*, *c*₁ and *c*₂ in higher plants, algae, and natural phytoplankton. *Biochem. Physiol. Pflanz.* **168**, 191–194 (1975).
- Ausubel, F. M. *et al.* *Short Protocols in Molecular Biology* (John Wiley & Sons, New York, 1992).
- Maidak, B. L. *et al.* The RDP (Ribosomal Database Project). *Nucleic Acids Res.* **25**, 109–111 (1997).
- Smith, S. W., Overbeek, R., Woese, C. R. & Gilbert, W. The genetic data environment: an expandable GUI for multiple sequence analysis. *CABIOS* **10**, 671–675 (1994).

Acknowledgements. We thank M. Sogin for access to an automated DNA sequencer and helpful discussion; A. Shimada for the use of unpublished sequences of *Prochlorococcus* GP2 and SB; R. Olson and E. Zettler for cell sorting at sea; H. Hsu for technical assistance; and C. Cavanaugh, H. Sosik and S. Sathyendranath for helpful consultation. This work was supported by the US National Science Foundation and the US National Aeronautics and Space Administration.

Correspondence and requests for materials should be addressed to S.W.C. (e-mail: chisholm@mit.edu). Sequences have been deposited in GenBank under accession numbers AF053396, MIT9302; AF053397, MIT9303; AF053398, MIT9312; and AF053399, MIT9313.

Conscious and unconscious emotional learning in the human amygdala

J. S. Morris*, A. Öhman† & R. J. Dolan*‡

* Wellcome Department of Cognitive Neurology, Queen Square, London WC1N 3BG, UK

† Department of Clinical Neuroscience, Section of Psychiatry and Psychology, Karolinska Hospital, S-17176 Stockholm, Sweden

‡ Royal Free and University College Hospitals School of Medicine, Rowland Hill Street, London NW3 2DF, UK

If subjects are shown an angry face as a target visual stimulus for less than forty milliseconds and are then immediately shown an expressionless mask, these subjects report seeing the mask but not the target. However, an aversively conditioned masked target can elicit an emotional response from subjects without being consciously perceived^{1,2}. Here we study the mechanism of this unconsciously mediated emotional learning. We measured neural activity in volunteer subjects who were presented with two angry faces, one of which, through previous classical conditioning, was associated with a burst of white noise. In half of the trials, the subjects' awareness of the angry faces was prevented by backward masking with a neutral face. A significant neural response was elicited in the right, but not left, amygdala to masked presentations of the conditioned angry face. Unmasked presentations of the same face produced enhanced neural activity

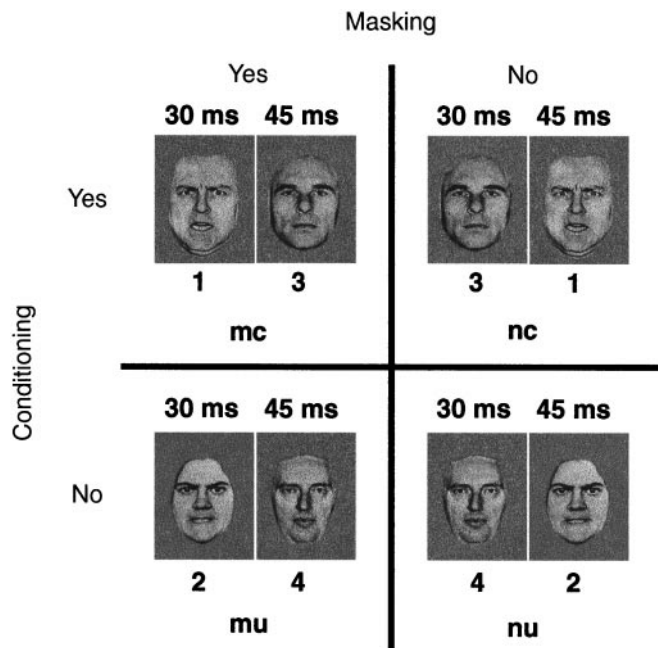


Figure 1 Stimulus parameters and experimental design. In the scanning window, pairs of target and masking faces were shown in four separate conditions, determined by the combination of masking and conditioning of the angry face. Mc, masked conditioned (the CS+ angry face was the target and the neutral face was the mask); nc, non-masked conditioned (a neutral face was the target and the CS+ face was the mask); mu, masked unconditioned (the CS– face was the target and the neutral face the mask); nu, non-masked unconditioned (the neutral face was the target and the CS– face the mask). Face 1, angry face paired with noise (CS+); face 2, angry face not paired with noise (CS–); faces 3 and 4, neutral faces. In all conditions, the target face was displayed for 30 ms and immediately followed by the mask for 45 ms.

in the left, but not right, amygdala. Our results indicate that, first, the human amygdala can discriminate between stimuli solely on the basis of their acquired behavioural significance, and second, this response is lateralized according to the subjects' level of awareness of the stimuli.

Studies of animals^{3–6} and brain-damaged patients^{7–10} indicate a crucial role for the amygdala in emotional learning. Other data show that the amygdala can respond differentially to emotional stimuli that subjects are unable explicitly to recall^{11,12}. However, direct evidence for the involvement of the amygdala in the unconscious mediation of learned emotional responses^{1,2} has been lacking. In this experiment we used a factorial design to measure directly how neural responses associated with emotional learning are modulated by subjects' conscious awareness (Fig. 1). We showed healthy volunteer subjects pictures of two angry faces, one of which was paired (CS+), and the other unpaired (CS–), with an unconditioned stimulus (UCS), namely, a 100-dB burst of white noise. Following this conditioning, the faces were presented sequentially, either masked or unmasked, and without the UCS, while neural activity was measured by positron emission tomography (PET). Throughout the experiment, subjects were required to indicate, by pressing a button, any occurrence of either angry face. Skin conductance responses (SCRs) were measured during both conditioning and scanning phases.

The responses of subjects revealed an inability to detect the masked angry faces: 0% of the masked angry faces were detected, whereas 100% of the unmasked faces were detected. On subsequent

debriefing, two subjects reported awareness of 'flickering' occurring on masked presentations, but did not report any perception of the angry faces. The remaining eight subjects denied any awareness of the masking procedure. Mean SCRs were significantly greater for CS+ than for CS– faces ($P < 0.01$), both for masked (CS+ mean = 0.605 μ S, s.e.m. = 0.06; CS– mean = 0.473, s.e.m. = 0.09 μ S) and unmasked (CS+ mean = 0.748 μ S, s.e.m. = 0.044; CS– mean = 0.426 μ S, s.e.m. = 0.084) presentations.

Bilateral responses in the amygdala were detected in scans of brain responses to all CS+ versus all CS– faces (masked and unmasked) (Table 1). To determine whether this response was present when subjects could not report the masked faces, we contrasted activity only in the masked CS+ and CS– scans. There was significant response ($P < 0.05$, corrected) in the region of the right amygdala to the presentation of masked CS+ faces (Fig. 2 and Table 1). This right-sided focus of activation was located in a medial and inferior part of the amygdaloid complex. In electrophysiological recordings of monkey brains, a similar region of amygdala was reported to contain units that are selective for faces and visually aversive cues¹³. When contrasting responses to unmasked CS+ and CS– faces, we saw a significant response ($P < 0.05$, corrected) in the left amygdala (Fig. 3 and Table 1). This left-sided activation by reportable CS+ faces was located in a more superior and posterior region of the amygdala than the right-sided activation by masked CS+ stimuli. Even with liberal thresholding, there was no evidence of responses of left amygdala to masked CS+ faces or of right amygdala to unmasked CS+ faces. Direct comparison of activity

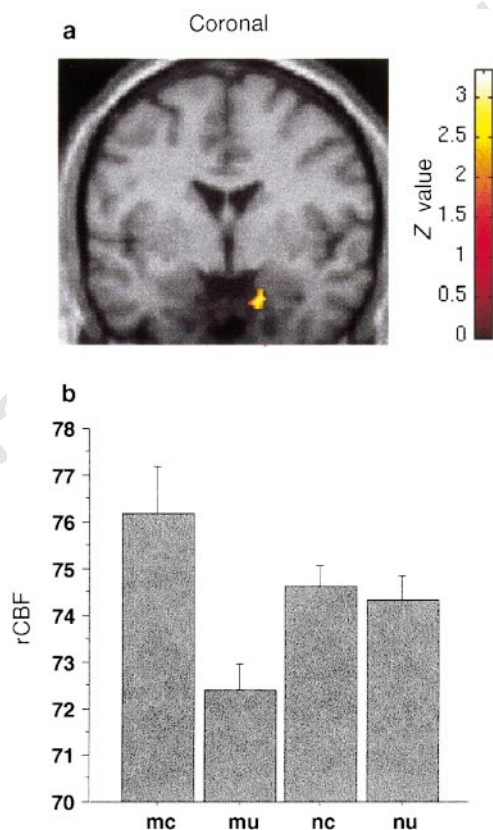


Figure 2 Response to the presentation of masked faces. **a**, An SPM showing activation in the region of the right amygdala in the contrast of masked CS+ and CS– angry faces. An uncorrected threshold of $P = 0.01$ was used to display the contrasts. The SPM is displayed on a coronal slice ($y = -2$ mm) of a canonical MRI image. **b**, A graphical display of the mean regional cerebral blood flow (rCBF) in the four conditions at the maximal voxel of activation in the right amygdala, $x = 18$, $y = -2$, $z = -28$. Bars represent two standard errors. See Fig. 1 legend for abbreviations.

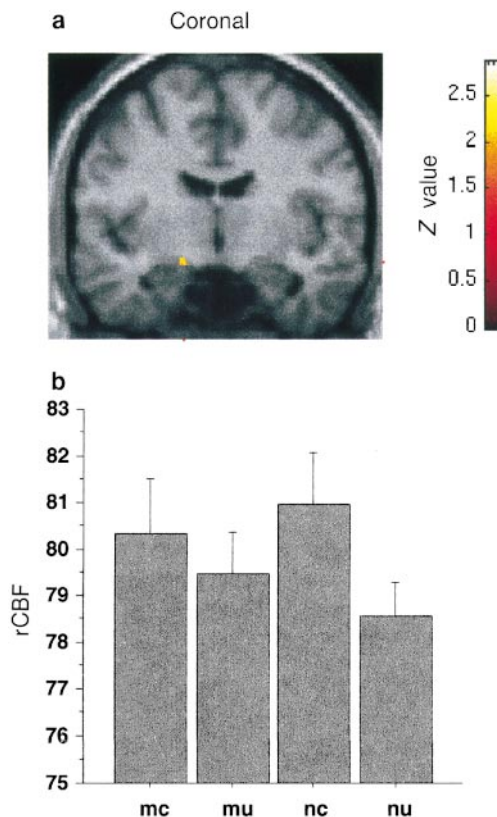


Figure 3 Response to the presentation of unmasked faces. **a**, A statistical parametric map (SPM) showing activation of the left amygdala in the contrast of unmasked CS+ and CS– angry faces. An uncorrected threshold of $P = 0.01$ was used to display the contrasts. The SPM is displayed on a coronal slice ($y = -8$ mm) of a canonical MRI image. **b**, A graphical display of the mean rCBF in the four conditions at the maximal voxel of activation in the left amygdala, $x = -18$, $y = -8$, $z = -14$. Bars represent two standard errors. See Fig. 1 legend for abbreviations.

Table 1 Brain regions showing significant activation in comparisons of responses to masked, unmasked, conditioned and non-conditioned faces

	Coordinates* (x, y, z)	Voxels*	Z score	P value (corrected)
All CS+ versus all CS–				
Left amygdala	–16, –10, –14	6	2.86	0.02
Right amygdala	18, –2, –28	21	2.67	0.04
Masked CS+ versus masked CS–				
Right amygdala	18, –2, –28	44	3.42	0.01
Unmasked CS+ versus unmasked CS–				
Left amygdala	–16, –8, –14	6	2.92	0.02

* Coordinates of the maximally activated voxel and the number of significant voxels are shown here.

between the two amygdalae revealed significant ($P < 0.05$) masking-related hemispheric effects.

We also formally tested the interaction between conditioning and masking, implicit in the above data, by comparing, for example, the responses to masked conditioned minus masked unconditioned faces with non-masked conditioned minus non-masked unconditioned faces. The results indicate that the response of the right amygdala to CS+ faces was significantly enhanced by masking (maximal voxel $x = 18$, $y = -2$, $z = -28$, $Z = 2.30$, $P < 0.05$), whereas the response of the left amygdala was enhanced by unmasking (maximal voxel $x = -14$, $y = -8$, $z = -14$, $Z = 1.65$, $P < 0.05$). As the different face stimuli were equally represented in conditions across subjects, and as no noise UCSs were played during scanning, the CS+ and CS– conditions involved identical physical stimuli and the same explicit task. The only difference, therefore, between the CS+ and CS– conditions was the subjects' previous experience of the temporal association between CS+ faces and aversive noise.

Our results are evidence that neural activity in the human amygdala mediates the learning of associations between behaviourally significant stimuli. The results concur with studies of fear conditioning in animals^{3–6} and of human lesions^{7–10} that also indicate that the amygdala is involved in emotional learning. The results are also consistent with neuroimaging data that show that amygdala activity during viewing of emotional stimuli correlates significantly with subsequent recall of the same stimuli¹⁴. A new aspect of our findings is that a response of the amygdala occurred to masked conditioned faces that were not consciously perceived. Although differential amygdala activity during passive viewing of masked fearful and happy faces has been reported¹², the subjects' awareness of the stimuli was not measured at the time of presentation, thus limiting interpretation of the data in terms of unconscious processing. In our study, however, we directly assessed subjects' awareness during scanning by an explicit target-detection task. Also, here the differential response of the amygdala was related not to differences in emotional expression, but to categorically identical stimuli that differed solely in terms of their earlier associative history (that is, one was associated with the aversive UCS). In this respect, we provide the first evidence that the human amygdala can discriminate the acquired behavioural significance of stimuli without the need for conscious perception.

Another new finding is the lateralization of the amygdala response as a function of the level of awareness of target stimuli. The response of the right amygdala to masked CS+ faces concurs with previous findings of a right-hemisphere advantage for processing emotional facial expressions^{15–17}. The absence of activity of the right amygdala in the unmasked condition, when subjects could report the occurrence of conditioned stimuli, indicates that processes related to conscious awareness, such as the engagement of language systems, may inhibit this neural response. This proposal is consistent with data from 'split-brain' patients which show that emotional visual stimuli presented to the right hemisphere produce greater autonomic responses when masked (and unreportable) than

when unmasked¹⁸. The proposal also concurs with electrophysiological data from humans concerning single-unit responses to previously presented stimuli; greater responses of the amygdala are produced in response to target stimuli that subjects could not recall¹¹. Also, psychophysical data show enhanced semantic priming with masked (unreportable) words; that is, there is attenuation of the priming effect by conscious awareness¹⁹. Our data may help to explain the failure of previous functional neuroimaging experiments using unmasked, reportable stimuli to show direct activation of the amygdala during conditioning paradigms^{20,21}. Subjects' conscious processing of stimuli in these studies may have attenuated amygdala responses.

The response of the right amygdala was attenuated in the unmasked condition, whereas activity of the left amygdala was enhanced with awareness of the stimuli. Left-hemisphere involvement in the recognition of facial expressions has been shown in several studies of human lesions^{22,23}. Previous functional imaging experiments have also reported left-lateralized responses in the amygdala that are related to the perception of facial expressions^{24,25}. All these studies used stimuli that subjects could identify by explicit report, indicating that the response of the left amygdala to emotional stimuli may be enhanced by verbal or conscious processing. The absence of a significant response of the left amygdala in the masked condition, when conscious (verbal) processing was prevented, supports this view. Neuroimaging data showing activation of right > left amygdala in response to masked fearful faces¹², but activation of left > right amygdala responses to the same stimuli without masking²⁵, are consistent with this interpretation.

Finally, the lateralization of the amygdala response here is consistent with the behavioural responses of patients with cerebral commissurotomies. 'Split-brain' patients can verbally report stimuli presented to the isolated left hemisphere yet deny awareness of the same stimuli presented to the right hemisphere²⁶. However, the isolated right hemisphere shows superior performance in tasks involving facial and emotional discrimination, when verbal processing is not critical^{16,26}. Greater heart-rate responses are provoked by masked emotional visual stimuli when presented to the right rather than left hemisphere of split-brain patients, but the same stimuli shown unmasked produce greater increases in heart rate after presentation to the isolated left hemisphere¹⁸. The lateralized amygdala activity here is consistent with these lesion data and indicates that the same segregation of conscious and unconscious processing that is observed in the split-brain may also be present in the intact brain. □

Methods

Ten healthy, right-handed male subjects took part in the study. Subjects (mean age 32.7 yr) were recruited by advertisement. They all gave informed consent and the study was approved by the local hospital ethics committee and ARSAC (UK). Each subject was scanned 12 times for the distribution of $H_2^{15}O$ using a Siemens/CPS ECAT EXACT HR + PET Scanner operated in high-sensitivity three-dimensional mode. Subjects received a total of 350 MBq of $H_2^{15}O$ over 20 s through a forearm cannula. Images were reconstructed into 63 planes, using a Hanning filter, resulting in 6.4-mm transaxial and 5.7-mm axial resolution (full width was half the maximum). Each scanning window was of duration 90 s.

PET data were analysed using statistical parametric mapping (SPM96) software from the Wellcome Department of Cognitive Neurology, London^{27,28}. After initial realignment, the PET scans were transformed into a standard stereotactic space²⁸. Structural magnetic resonance images (MRIs) from each subject were co-registered into the same space. The scans were then smoothed using a Gaussian filter, set at 12 mm full width at half maximum. We adjusted the rCBF (regional cerebral blood flow) measurements to a global mean of 50 ml per dl per min. A blocked (by subject) analysis of covariance model was fitted to the data at each voxel, with condition effects for each of the four masking variations, and global CBF as a confounding covariate. We assessed predetermined contrasts of the condition effects at each voxel using a *t*-statistic,

giving a statistic image for each contrast. *P* values for activations in the amygdala were corrected for the volume of brain analysed (specified as a sphere with radius 8 mm)²⁹. Anatomical localization for the group mean-condition-specific activations are reported in standard space²⁸. In all cases, the localization of the group mean activations was confirmed by registration with the subject's own MRIs.

In an initial conditioning phase immediately before scanning, subjects viewed a sequence of greyscale images of four faces taken from a standard set of pictures of facial affect³⁰. Images of a single face were presented on a computer monitor screen for 75 ms at intervals of 15–25 s (mean 20 s). Each of the four faces was shown six times in a pseudorandom order. Two of the faces had angry expressions (A1 and A2), the other two being neutral (N1 and N2). One of the angry faces (CS+) was always followed by a 1-s 100-dB burst of white noise. In half of the subjects A1 was the CS+ face; in the other half, A2 was used. None of the other faces was ever paired with the noise. Before each of the 12 scanning windows, which occurred at 8-min intervals, a shortened conditioning sequence was played consisting of three repetitions of the four faces. During the 90-s scanning window, which seamlessly followed the conditioning phase, 12 pairs of faces, consisting of a target and mask, were shown at 5-s intervals. The target face was presented for 30 ms and was immediately followed by the masking face for 45 ms (Fig. 1). These stimulus parameters remained constant throughout all scans and effectively prevented any reportable awareness of the target face (which might be a neutral face or an angry face).

There were four different conditions (Fig. 1), masked conditioned, non-masked conditioned, masked unconditioned, and non-masked unconditioned. Throughout the experiment, subjects performed the same explicit task, which was to detect any occurrence, however fleeting, of the angry faces. Immediately before the first conditioning sequence, subjects were shown the two angry faces and were instructed, for each stimulus presentation, to press a response button with the index finger of the right hand if one the angry faces appeared, or another button with the middle finger of the right hand if they did not see either of the angry faces.

Throughout the acquisition and extinction phases, subjects' SCRs were monitored to index autonomic conditioning. SCRs were measured with Biodata galvanic skin response equipment using Ag/AgCl electrodes attached to the palmar surface of the middle phalanges of the index and middle fingers of the left hand. We took readings of skin conductance (in μ S) every 500 ms and stored them digitally on computer. All SCRs were square-root-transformed to attain statistical normality. Using the SCR in the 4-s period before presentation as a baseline, the maximal SCR deflection in the period 0.5–4 s after a face was presented was assigned as the value for the SCR to that face. The mean SCRs for the CS+ and CS- angry faces were calculated for both the masked and the unmasked conditions, and the differences between the means were tested using a paired Student's *t*-test.

Received 2 February; accepted 13 March 1998.

- Estevés, F., Dimberg, U. & Ohman, A. Automatically elicited fear: conditioned skin conductance responses to masked facial expressions. *Cogn. Emotion* **9**, 99–108 (1994).
- Ohman, A. & Soares, J. J. F. "Unconscious anxiety": phobic responses to masked stimuli. *J. Abnorm. Psychol.* **103**, 231–240 (1994).
- Weiskrantz, L. Behavioural changes associated with ablation of the amygdaloid complex in monkeys. *J. Comp. Physiol. Psychol.* **49**, 381–391 (1956).
- Kapp, B. S., Pascoe, J. P. & Bixler, M. A. in *Neuropsychology of Memory* (eds Butters, N. & Squire, L. R.) 473–488 (Guilford, New York, 1984).
- Davis, M. in *The Amygdala: Neurobiological Aspects of Emotion, Memory and Mental Dysfunction* (ed. Aggleton, J. P.) 255–306 (Wiley-Liss, New York, 1992).
- LeDoux, J. E. *The Emotional Brain* (Simon and Schuster, New York, 1996).
- Bechara, A. et al. Double dissociation of conditioning and declarative knowledge relative to the amygdala and hippocampus in humans. *Science* **269**, 1115–1118 (1995).
- LaBar, K. S., LeDoux, J. E., Spencer, D. D. & Phelps, E. A. Impaired fear conditioning following unilateral temporal lobectomy. *J. Neurosci.* **15**, 6846–6855 (1995).
- Babinsky, R. et al. The possible contribution of the amygdala to memory. *Behav. Neurol.* **6**, 167–170 (1993).
- Adolphs, R., Cahill, L., Scull, R. & Babinsky, R. Impaired declarative memory for emotional material following bilateral amygdala damage in humans. *Learning Memory* **4**, 291–300 (1997).
- Fried, I., MacDonald, K. A. & Wilson, C. L. Single neuron activity in hippocampus and amygdala during recognition of faces and objects. *Neuron* **18**, 875–887 (1997).
- Whalen, P. J. et al. Masked presentations of emotional facial expressions modulate amygdala activity without explicit knowledge. *J. Neurosci.* **18**, 411–418 (1998).
- Leonard, C. M., Rolls, E. T., Wilson, F. A. & Baylis, G. C. Neurons in the amygdala of the monkey with responses selective for faces. *Behav. Brain Res.* **15**, 159–176 (1985).
- Cahill, L. et al. Amygdala activity at encoding correlated with long-term, free recall of emotional information. *Proc. Natl Acad. Sci. USA* **93**, 8016–8021 (1996).
- DeKosky, S. T., Heilman, K. M., Bowers, D. & Valenstein, E. Recognition and discrimination of emotional faces and pictures. *Brain Lang.* **9**, 206–214 (1980).

- Gazzaniga, M. S. & Smylie, C. S. Facial recognition and brain asymmetries: clues to underlying mechanisms. *Ann. Neurol.* **13**, 536–540 (1983).
- Gur, R. C., Skolnick, B. E. & Gur, R. E. Effects of emotional discrimination tasks on cerebral blood flow: regional activation and its relation to performance. *Brain Cogn.* **25**, 271–286 (1994).
- Ladavas, E., Cimatti, D., Del Pesce, M. & Tuozi, G. Emotional evaluation with and without conscious stimulus identification: evidence from a split-brain patient. *Cogn. Emotion* **7**, 95–114 (1993).
- Klinger, M. R. & Greenwald, A. G. Unconscious priming of association judgements. *J. Exp. Psych. Learn. Mem. Cogn.* **21**, 569–581 (1995).
- Hugdahl, K. et al. Brain mechanisms in human classical conditioning: a PET blood flow study. *Neuroreport* **6**, 1723–1728 (1995).
- Morris, J. S., Friston, K. J. & Dolan, R. J. Neural responses to salient visual stimuli. *Proc. R. Soc. Lond. B* **264**, 769–775 (1997).
- Young, A. W., Newcombe, F., de Haan, E. H. F., Small, M. & Hay, D. C. Face perception after brain injury. Selective impairments affecting identity and expression. *Brain* **116**, 941–959 (1993).
- Stone, V. S., Nisenson, L., Eliassen, J. C. & Gazzaniga, M. S. Left hemisphere representations of emotional facial expressions. *Neuropsychol.* **34**, 23–29 (1995).
- Morris, J. S. et al. A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature* **383**, 812–815 (1996).
- Breiter, H. C. et al. Response and habituation of the human amygdala during visual processing of facial expression. *Neuron* **17**, 875–887 (1996).
- Gazzaniga, M. S. in *The Cognitive Neurosciences* (ed. Gazzaniga, M. S.) 1049–1061 (MIT Press, Cambridge, MA, 1995).
- Friston, K. J. et al. Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* **2**, 189–210 (1995).
- Friston, K. J. et al. Spatial registration and normalization of images. *Hum. Brain Mapp.* **3**, 165–189 (1995).
- Worsley, K. J., Marrett, P., Neelin, A. C., Friston, K. J. & Evans, A. C. A unified statistical approach for determining significant signals in images of cerebral activation. *Hum. Brain Mapp.* **4**, 58–73 (1996).
- Ekman, P. & Friesen, W. V. *Pictures of Facial Affect* (Consulting Psychologists Press, Palo Alto, 1976).

The human amygdala in social judgment

Ralph Adolphs*, Daniel Tranel* & Antonio R. Damasio*†

* Department of Neurology, Division of Cognitive Neuroscience, University of Iowa College of Medicine, 200 Hawkins Drive, Iowa City, Iowa 52242, USA

† The Salk Institute for Biological Studies, La Jolla, California 92037, USA

Studies in animals have implicated the amygdala in emotional^{1–3} and social^{4–6} behaviours, especially those related to fear and aggression. Although lesion^{7–10} and functional imaging^{11–13} studies in humans have demonstrated the amygdala's participation in recognizing emotional facial expressions, its role in human social behaviour has remained unclear. We report here our investigation into the hypothesis that the human amygdala is required for accurate social judgments of other individuals on the basis of their facial appearance. We asked three subjects with complete bilateral amygdala damage to judge faces of unfamiliar people with respect to two attributes important in real-life social encounters: approachability and trustworthiness. All three subjects judged unfamiliar individuals to be more approachable and more trustworthy than did control subjects. The impairment was most striking for faces to which normal subjects assign the most negative ratings: unapproachable and untrustworthy looking individuals. Additional investigations revealed that the impairment does not extend to judging verbal descriptions of people. The amygdala appears to be an important component of the neural systems that help retrieve socially relevant knowledge on the basis of facial appearance.

Data from three subjects with complete bilateral amygdala damage (subjects SM, JM and RH) and seven with unilateral amygdala damage were compared to those from normal and from brain-damaged control subjects (see Table 1 and Methods). Ratings of approachability and of trustworthiness were analysed separately for the 50 faces to which normal controls assigned the most negative ratings, and for the 50 most positive faces. Subjects with bilateral amygdala damage rated the 50 most negative faces more positively than did either normal controls ($P < 0.01$) or brain-damaged controls ($P < 0.05$; Mann–Whitney *U*-tests on subjects' mean ratings, Bonferroni corrected) (Fig. 1). Groups with unilateral amygdala lesions did not differ from controls on either rating. All subject groups gave similar ratings to the 50 most positive faces.

Subject SM spontaneously commented during the experiment that, in real life, she would not know how to judge if a person were trustworthy, consistent with her tendency to approach and engage in physical contact with other people rather indiscriminately^{7,14}. All subjects with bilateral amygdala damage had normal ability to discriminate faces (Table 1), clear evidence that there were no visuo-perceptual impairments that might account for the above findings.

Data from subjects with bilateral amygdala damage showed two effects: the subjects tended to rate all faces more positively than did controls, and they also showed the largest deviation from control ratings specifically when rating the most negative faces (Fig. 2). This suggests an overall positive bias, as well as a disproportionate impairment in rating the most negative faces. To establish the independence of these two effects, we carried out a detailed two-alternative forced-choice task with JM, RH and SM, using the same 100 face stimuli. We asked JM and RH to choose the more approachable face in pairwise comparisons between an anchor face that received a mean normal rating of 0.0 and each of the remaining 99 faces. We compared subjects' choices on this task to the choices that would be expected from the mean approachability ratings given to the faces in each pair by normal controls. JM and RH consistently made more incorrect choices when making comparisons to very negative faces, than when making comparisons to very positive faces (Fig. 3a). By contrast, the small number of errors made by normal controls occurred in the opposite direction, with positive rather than with negative faces (Fig. 3a), indicating that the impairments seen in amygdala subjects cannot be explained by stimulus difficulty.

In subject SM, we carried out forced-choice tasks with a total of five anchor faces, including faces normally rated very negatively and very positively. Each anchor face was paired with the remaining 99 faces, for a total of $5 \times 99 = 495$ pairwise comparisons. SM made the largest number of incorrect choices in comparisons involving those of the five anchor faces that normally receive the most negative

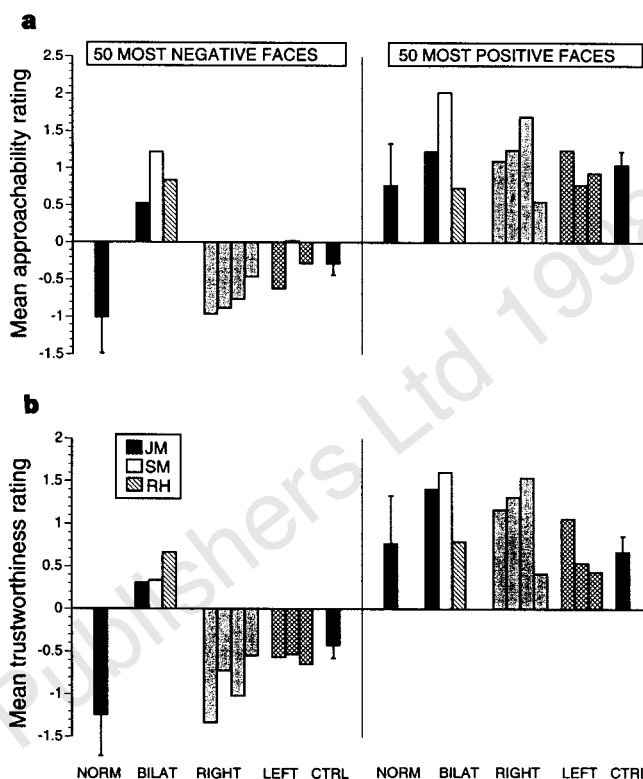


Figure 1 Mean judgments. **a**, Approachability; **b**, trustworthiness of the faces of 100 unfamiliar people, shown for the 50 faces that received the most negative (left) and most positive (right) mean ratings from normal controls. Data are shown from 46 normal controls (NORM; means and s.d.), 3 subjects with bilateral amygdala damage (BILAT; individual means), 4 subjects with unilateral right (RIGHT) and 3 with unilateral left (LEFT) amygdala damage, and 10 brain-damaged controls with no damage to amygdala (CTRL; means and s.e.m.).

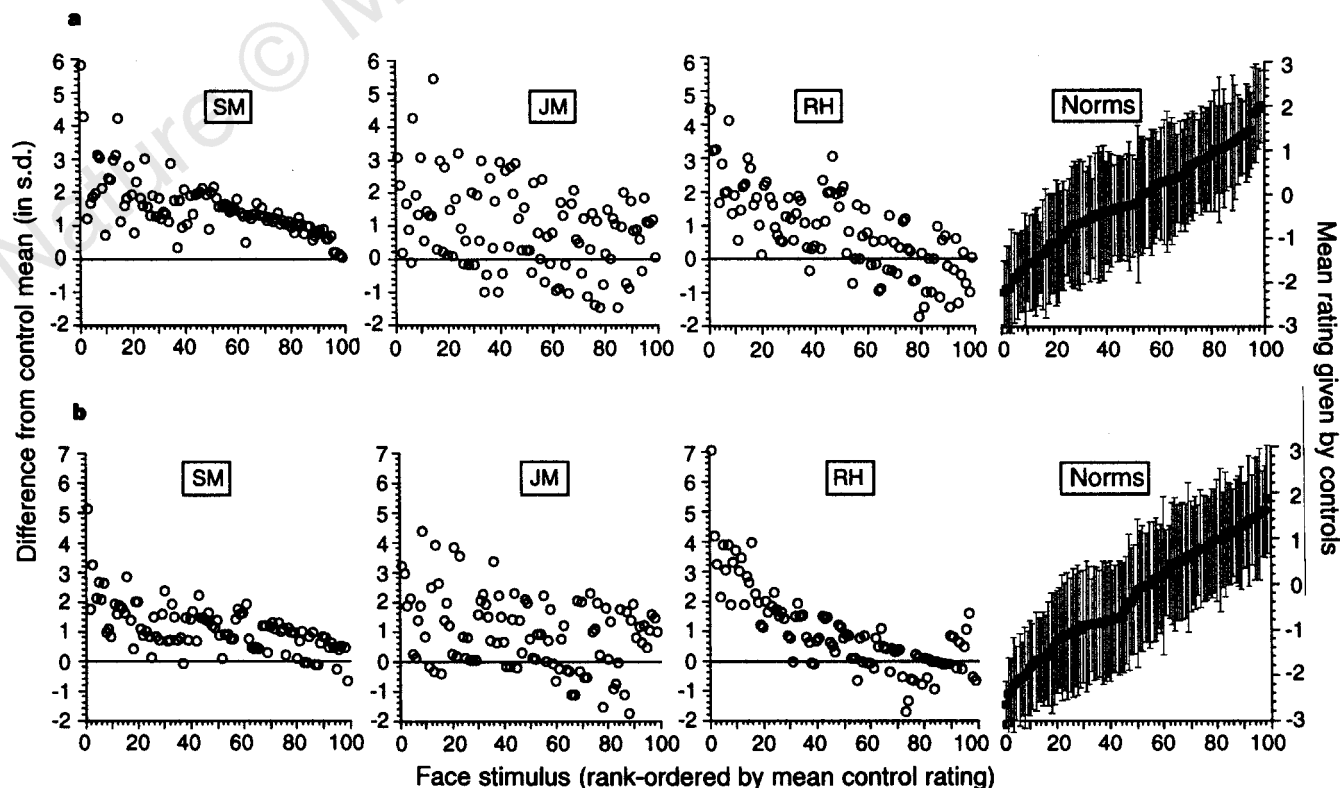


Figure 2 Deviations from normal judgments. **a**, Approachability; **b**, trustworthiness given by subjects with bilateral amygdala damage (circles; left y-axis). Units are standard deviations of the normal control ratings. Stimuli are rank-ordered on

the x-axis according to the ratings normal controls gave them (squares; far right y-axis; means and s.d.).

ratings, a performance that was highly abnormal compared to control subjects' forced-choices in the same task (Fig. 3b). The findings from the forced-choice tasks cannot be explained solely on the basis of a general positive bias, and confirm that judgments given by subjects with bilateral amygdala damage are disproportionately impaired relative to individuals who are normally classified as unapproachable.

Might the impairment seen in subjects with bilateral amygdala damage extend to judging people from word descriptions rather than from faces? We asked subjects to rate the likeability of different individuals based on short verbal biographies or on single words (adjectives describing people). All three subjects with bilateral amygdala damage made entirely normal judgments when the stimuli were verbal (Fig. 4). This critical dissociation supports the

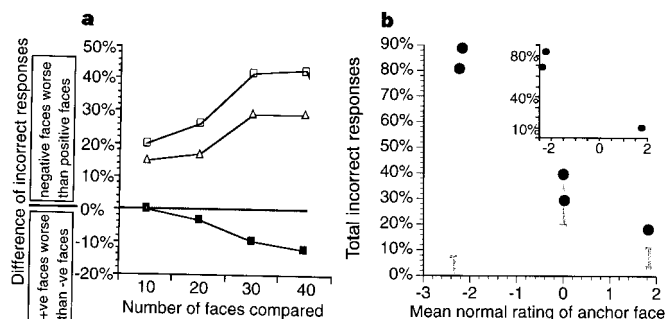


Figure 3 Disproportionate impairment in choosing the most unapproachable faces. **a**, JM's (empty squares), RH's (triangles) and normal controls' (filled squares) judgments of approachability from two-alternative forced-choice tasks. We calculated the per cent incorrect choices made for pairings involving faces at either extreme of the normal rating scale (that is, faces that were normally rated as either very approachable or very unapproachable). The x-axis shows the number of faces at either extreme of the normal rating scale over which the per cent incorrect choices was calculated. The y-axis shows the difference in the errors made (unapproachable – approachable). **b**, SM's judgments of approachability from two-alternative forced-choice tasks. The mean normal rating of approachability given to each of 5 anchor faces is shown on the x-axis, and the proportion of incorrect forced choices (out of 99) made by SM (circles) and by normal controls (grey bars show range) are shown on the y-axis. Inset, analysis of SM's data from this task for only those pairs of faces whose mean control rating differed by more than 2 rating points. Data from comparisons involving the two faces with mean control ratings of 0 were not analysed, as very few faces with ratings < -2 or > 2 could be paired with them. No normal control made any incorrect choices in this analysis.

following interpretation. The amygdala appears necessary to trigger the retrieval of information on the basis of prior social experience or innate bias in regard to certain classes of faces¹⁵. The retrieved information might be either covert or overt, or both (compare with ref. 16). The failure due to amygdala damage thus occurs after basic visual processing has taken place, by blocking the retrieval of information normally linked either to negative past experiences with similar stimuli, or to innately specified feature configurations. By contrast, sentences and words evoke a broad sweep of information directly, without the need for the amygdala's assistance, thus providing a sufficient basis for performing judgments normally.

A further question concerns the specific facial cues that would normally engage the amygdala in social judgment. Might amygdala lesions impair judgments based only on certain facial features? This does not seem to be the case, as subjects with bilateral amygdala lesions gave idiosyncratic ratings to specific negative faces (Fig. 2; intersubject Spearman rank correlations of ratings given by subjects with bilateral amygdala lesions for the 50 most negative faces: $-0.23 < r < 0.31$). We further explored this complex issue by choosing the 10 faces to which SM had given the most abnormal ratings of approachability (all rated very negatively by controls), and systematically manipulating individual features in each face. We showed subjects 109 pairs of faces in which each pair showed the same individual differing by only one single feature. We manipulated direction of gaze (45 stimuli), expression of the eyes (27 stimuli), expression of the mouth (14 stimuli), or visibility of the eyes (for example, with glasses of different tint; 23 stimuli), all features that might conceivably contribute to the subjects' judgments. In a two-alternative forced-choice task, SM and 16 normal controls were asked to choose the face they would prefer to approach. SM performed entirely normally on this task. Logistic linear analysis, with subjects' binary choices as the dependent variable and the manipulated features as factors, showed that SM did not differ from controls in her choices with respect to any of the above features that we had manipulated. Insensitivity to particular features, in isolation, is thus unlikely to account for the impairment in judging approachability or trustworthiness in faces.

The findings suggest that the human amygdala triggers socially and emotionally relevant information in response to visual stimuli. The amygdala's role appears to be of special importance for social judgment of faces that are normally classified as unapproachable and untrustworthy, consistent with the amygdala's demonstrated role in processing threatening and aversive stimuli. An intriguing question that remains to be addressed is the amygdala's relative participation in triggering information that is innate, versus infor-

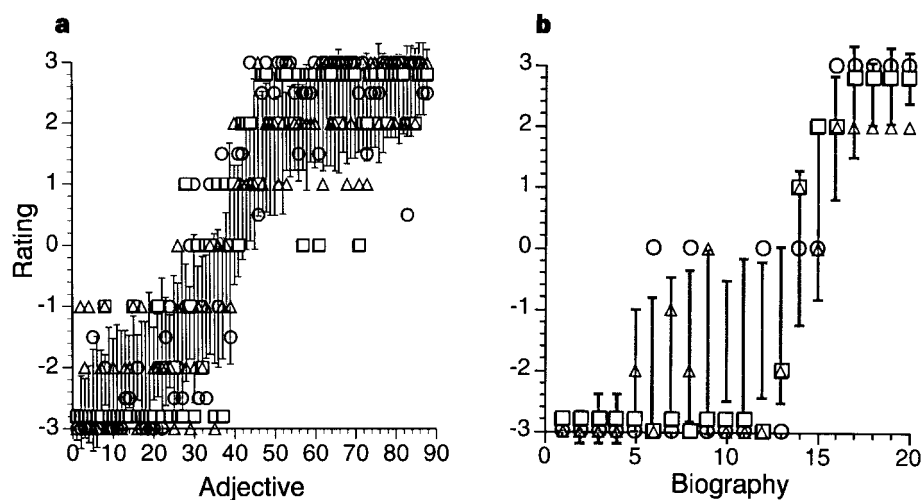


Figure 4 Likeability ratings of lexical stimuli. **a**, Ratings given by SM (2 experiments; circles), JM (squares), RH (triangles) and 20 normal controls (s.d.).

shown as bars) to 88 adjectives describing personality. **b**, Ratings given by SM, JM, RH and 20 normal controls to 20 biographical descriptions of people.

Table 1 Neuropsychology background data

Subject	SM	JM	RH	Left	Right	BD Ctrl	N Ctrl
N				3	4	10	46
Age	31	67	42	29 ± 5	31 ± 5	66 ± 10	19 ± 1
PIQ	90	95	108	104 ± 24	103 ± 24	98 ± 11	-
Benton (%ile)	71st	12th	20th	32nd	50th	48th	-
Expression discrimination	70 (40)	45 (20)	59 (15)	-	-	-	-
Gender discrimination	75th	65th	65th	-	-	-	-
Gaze discrimination	64th	86th	86th	-	-	-	-

SM, JM, RH, bilateral amygdala damage; Left, Right, unilateral amygdala damage; BD Ctrl and N Ctrl, brain-damaged and normal controls. PIQ, performance IQ from the Wechsler Adult Intelligence Scale-revised. Benton, percentile score on the Benton Facial Recognition Test, a measure of ability to discriminate among unfamiliar faces¹⁸. Expression, gender, gaze discrimination, percentile score on two-alternative forced-choice discrimination tasks of emotional facial expressions (average and minimum (in parentheses) for the 6 basic emotions), gender, and direction of gaze. See Methods for details.

mation that is acquired through individual experience in a cultural setting¹⁷. □

Methods

Subjects. All subjects had given informed consent to participate in these studies. Brain-damaged subjects were selected from the Patient Registry of the Department of Neurology at the University of Iowa, and had been fully characterized neuropsychologically¹⁸ and neuroanatomically^{19,20}.

Amygdala damage. Subject SM has complete lesions of both amygdalae, as well as minimal damage to anterior entorhinal cortex, resulting from Urbach-Wiethe disease^{7,14,21,22}. Subjects RH and JM had encephalitis at ages 28 and 62, respectively, resulting in complete bilateral destruction of the amygdala and substantial damage to surrounding structures. Both patients are severely amnesic. Seven subjects with unilateral amygdala lesions (4 right, 3 left) had surgical temporal lobectomy for the treatment of epilepsy, and also had damage to hippocampus and surrounding temporal cortices.

Control subjects. We examined 10 brain-damaged controls with lesions that did not include the amygdala. Four of the subjects had bilateral lesions. Three of the subjects were amnesic consequent to anoxia and hippocampal damage. We also examined 46 normal controls (16M/30F) who were undergraduates at the University of Iowa.

Stimuli and tasks. In all tasks, stimuli within each session were presented in randomized order, and without time limit.

Approachability and trustworthiness ratings of faces. We selected from a larger set of photographs 100 final stimuli whose ratings had low variance and were evenly distributed (Fig. 2). There was no effect of subject gender on rating the faces ($P > 0.7$, ANOVA on normal data). Stimuli were black-and-white photographs of unfamiliar male ($N = 55$) and female ($N = 45$) faces in natural poses.

Subjects were asked to rate the stimuli, shown one at a time on a slide projector, on a 7-point scale (-3 to +3) with respect to either approachability or trustworthiness. For approachability, subjects were asked to imagine meeting the person on the street, and to indicate how much they would want to walk up to that person and strike up a conversation. For trustworthiness, subjects imagined trusting that person with all their money, or with their life. Each of the two attributes was rated in two independent sessions in counterbalanced order; there were no order effects.

Approachability and trustworthiness were chosen because (1) they are clear measures of real-life social judgment; (2) they are easy to understand; and (3) pilot data indicated that ratings of these specific attributes had lower variance than those obtained with other words, such as 'nice' or 'good'. Although approachability and trustworthiness ratings in normals were somewhat correlated (mean $r = 0.52$), there were many stimuli that received discrepant ratings on the two attributes, indicating that they were non-redundant measures of social judgment.

Forced-choice tasks. Direct pairwise comparisons of approachability were made between an anchor face, and each of the remaining 99 faces, all drawn from the same 100 face stimuli used in other tasks. We calculated the proportion of subjects' choices that differed from the choices that would be expected on the basis of the mean normal control ratings given to each of the two faces in a pair.

In one experiment (Fig. 3a), each of two anchor faces with a mean normal approachability rating of 0.0 was compared to other faces that were either very

approachable or very unapproachable; data obtained with both anchor faces were very similar and were pooled. In a second experiment (Fig. 3b), each of 5 anchor faces (which included faces with a range of ratings) was compared to all other 99 faces (a total of 495 pairwise comparisons).

Lexical stimuli. We chose 88 adjectives that described personality attributes from a large standardized set²³ so as to span the range from very likeable to very dislikeable, and to exhibit maximal reliability and common usage. Twenty short biographies described people by giving information about the person's lifestyles and activities. Subjects rated how much they liked individuals described by the stimuli, on a scale of -3 to +3. Words were presented visually on a sheet of paper; biographical descriptions were read to subjects.

Control tasks. For each of the control tasks, we calculated thresholds at which subjects were just able to discriminate stimuli. Data were converted to percentiles compared to performances given by normal subjects ($N = 28$ for expression, 20 for gender, 28 for gaze).

Expression discrimination. Two-alternative forced-choice discriminations were made between 80 images of a neutral face, and 80 images that were linear morphs between the neutral face and facial expressions of emotion²⁴ (happiness, surprise, fear, anger, disgust, sadness). Subjects were asked to choose the image that showed more of a stated emotion.

Gender discrimination. Two-alternative forced-choice discriminations were made between 84 pairs of images that were morphs between an average composite of a neutral male face, and an average composite of a neutral female face, of equal age.

Gaze discrimination. Two-alternative forced-choice discriminations were made between 16 pairs of images showing the same, neutral, male face in which only direction of gaze had been varied by manipulating the digital image on a computer.

Received 7 January; accepted 30 March 1998.

- Weiskrantz, L. Behavioral changes associated with ablation of the amygdaloid complex in monkeys. *J. Comp. Physiol. Psychol.* **49**, 381-391 (1956).
- Blanchard, D. C. & Blanchard, R. J. Innate and conditioned reactions to threat in rats with amygdaloid lesions. *J. Comp. Physiol. Psychol.* **81**, 281-290 (1972).
- Le Douarin, J. *The Emotional Brain* (Simon and Schuster, New York, 1996).
- Rosvold, H. E., Mirsky, A. F. & Pribram, K. Influence of amygdalotomy on social behavior in monkeys. *J. Comp. Physiol. Psychol.* **47**, 173-178 (1954).
- Kling, A., Steklis, H. D. & Deutsch, S. Radiotelemetered activity from the amygdala during social interactions in the monkey. *Exp. Neurol.* **66**, 88-96 (1979).
- Kling, A. S. & Brothers, L. A. in *The Amygdala: Neurobiological Aspects of Emotion, Memory, and Mental Dysfunction* (ed. Aggleton, J. P.) 353-378 (Wiley-Liss, New York, 1992).
- Adolphs, R., Tranel, D., Damasio, H. & Damasio, A. Impaired recognition of emotion in facial expressions following bilateral damage to the human amygdala. *Nature* **372**, 669-672 (1994).
- Young, A. W. et al. Face processing impairments after amygdalotomy. *Brain* **118**, 15-24 (1995).
- Calder, A. J. et al. Facial emotion recognition after bilateral amygdala damage: differentially severe impairment of fear. *Cogn. Neuropsychol.* **13**, 699-745 (1996).
- Broks, P. et al. Face processing impairments after encephalitis: amygdala damage and recognition of fear. *Neuropsychologia* **39**, 59-70 (1998).
- Morris, J. S. et al. A differential neural response in the human amygdala to fearful and happy facial expressions. *Nature* **383**, 812-815 (1996).
- Breiter, H. C. et al. Response and habituation of the human amygdala during visual processing of facial expression. *Neuron* **17**, 875-887 (1996).
- Morris, J. S. et al. A neuromodulatory role for the human amygdala in processing emotional facial expressions. *Brain* **121**, 47-57 (1998).
- Tranel, D. & Hyman, B. T. Neuropsychological correlates of bilateral amygdala damage. *Arch. Neurol.* **47**, 349-355 (1990).
- Damasio, A. R. Toward a neurobiology of emotion and feeling: operational concepts and hypotheses. *Neuroscientist* **1**, 19-25 (1995).
- Lewicki, P., Hill, T. & Czerwinski, M. Nonconscious acquisition of information. *Am. Psychol.* **47**, 796-801 (1992).
- Saarni, C., Mumme, D. L. & Campos, J. J. in *Handbook of Child Psychology*, Vol. 3: *Social, Emotional, and Personality Development* (ed. Damon, W.) 237-309 (Wiley, New York, 1997).

18. Tranel, D. in *Neuropsychological Assessment of Neuropsychiatric Disorders* (eds Grant, I. & Adams, K. M.) 81–101 (Oxford Univ. Press, New York, 1996).
19. Damasio, H. & Frank, R. Three-dimensional *in vivo* mapping of brain lesions in humans. *Arch. Neurol.* **49**, 137–143 (1992).
20. Frank, R. J., Damasio, H. & Grabowski, T. J. Brainvox: an interactive, multi-modal visualization and analysis system for neuroanatomical imaging. *NeuroImage* **5**, 13–30 (1997).
21. Adolphs, R., Tranel, D., Damasio, H. & Damasio, A. R. Fear and the human amygdala. *J. Neurosci.* **15**, 5879–5892 (1995).
22. Nahm, F. K. D., Tranel, D., Damasio, H. & Damasio, A. R. Cross-modal associations and the human amygdala. *Neuropsychologia* **31**, 727–744 (1993).
23. Anderson, N. H. Likableness ratings of 555 personality-trait words. *J. Person. Social Psychol.* **9**, 272–279 (1968).
24. Ekman, P. & Friesen, W. *Pictures of Facial Affect* (Consulting Psychologists, Palo Alto, CA, 1976).

Acknowledgements. We thank J. Suhr and J. Nath for technical assistance in testing subjects, D. Krutzfeldt for help in scheduling subjects and H. Damasio for comments on the manuscript. This study was supported by a grant from the National Institute for Neurological Diseases and Stroke.

Correspondence and requests for materials should be addressed to R.A. (e-mail: ralph-adolphs@uiowa.edu).

A conditioned dendritic cell can be a temporal bridge between a CD4⁺ T-helper and a T-killer cell

John Paul Ridge*, Francesca Di Rosa*† & Polly Matzinger*

* Ghost lab, Section on T cell Tolerance and Memory, Laboratory of Cellular and Molecular Immunology, NIAID/NIH Building 4 Room 111, 9000 Rockville Pike, Bethesda, Maryland 20892, USA

To generate an immune response, antigen-specific T-helper and T-killer cells must find each other and, because they cannot detect each other's presence, they are brought together by an antigen-loaded dendritic cell that displays antigens to both^{1–3}. This three-cell interaction, however, seems nearly impossible because all three cell types are rare and migratory. Here we provide a potential solution to this conundrum. We found that the three cells need not meet simultaneously but that the helper cell can first engage and 'condition' the dendritic cell, which then becomes empowered to stimulate a killer cell. The first step (help) can be bypassed by modulation of the surface molecule CD40, or by viral infection of dendritic cells. These results may explain the long-standing paradoxical observation that responses to some viruses are helper-independent, and they evoke the possibility that dendritic cells may take on different functions in response to different conditioning signals.

We began our study to discriminate between two interpretations of this three-cell interaction (Fig. 1). The antigen-presenting cell (APC) has been proposed to have a rather passive relationship with the killer cell (also known as a cytotoxic T lymphocyte) and to function mainly to stimulate the helper cell to produce the interleukin (IL)-2 that the killer needs^{1,2} (Fig. 1a). There is no guarantee, however, that a rare helper and an equally rare killer should find the same APC at the same time. As resting killers recognizing antigen become tolerant if there is no help^{4–6}, many potentially useful killers would founder while, elsewhere, some T helpers would wastefully secrete cytokines into an environment containing no killers to receive them. We therefore suggested a dynamic model (Fig. 1b) in which the T helper stimulates the APC to become able to activate the killer⁶.

To discriminate between these possibilities, we studied responses to the male antigen H–Y because, first, killers that recognize H–Y are helper-dependent^{6,7}; second, H–Y has no known crossreactive environmental mimics⁸; and third, primary and secondary

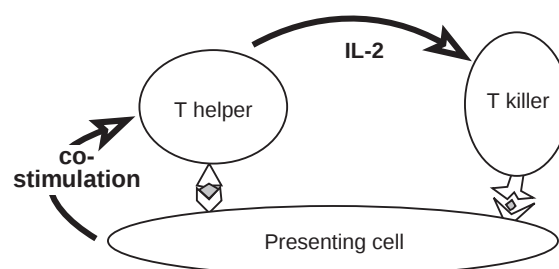
responses can be easily distinguished because T cells from normal virgin female mice respond *in vitro* only if they were first primed *in vivo* with professional APCs^{7,9}.

Figure 2a–c shows that help is necessary for generation of killing activity against H–Y and that it can be replaced by soluble factors¹. Female C57Bl/6 (B6) mice, immunized *in vivo* with male spleen cells, generated good *in vitro* killer-cell responses against male spleen stimulators (Fig. 2a). The responses disappeared if we removed the CD4⁺ cells just before the culture (Fig. 2b) and reappeared if we added soluble helper factors (concanavalin A supernatant (CAS); Fig. 2c).

In some cases where help is minimal, such as in newborns (which have very few T cells) and in B6.bm12 mice (with mutated major histocompatibility complex (MHC) class II molecules), a killer-cell response can be induced by an injection of activated male dendritic cells^{10,11}. We found, however, that activated dendritic cells could not stimulate purified CD8⁺ killer cells unless we added helper cells, in the form of Marilyn, an H–Y-specific T-helper clone (Fig. 2e). Thus a small number of helpers may go a long way but without them dendritic cells are unable to activate killers against H–Y.

Because activated T helpers express CD40-ligand, which can stimulate CD40 to induce proliferation in B cells¹² and enhance the function of dendritic cells^{13,14}, we tried replacing T-cell help with antibodies against CD40. We found that overnight crosslinking with anti-CD40 antibodies turned dendritic cells into excellent stimulators (Fig. 2f). To rule out the possibility that the crosslinked dendritic cells were simply stimulating better IL-2 production from a few contaminating CD4⁺ cells, we tested dendritic cells from MHC-class II-knockout (MHC II KO) mice, which are deficient in MHC class II molecules because of a gene-targeted deletion. Although these dendritic cells cannot present antigen to CD4⁺ helpers, they became good stimulators for killers (Fig. 2h).

a Three-cell interaction



b Sequential two-cell interactions

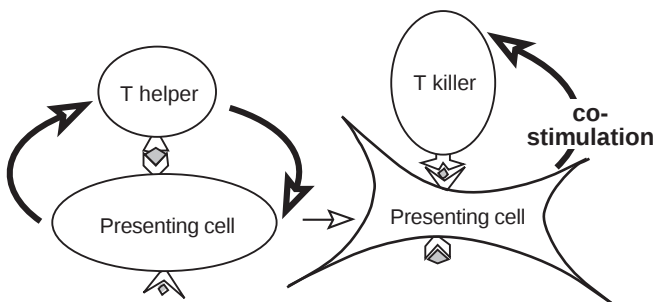


Figure 1 Two models of the delivery of help to CD8⁺ killers. **a**, The 'passive' model in which the dendritic (presenting) cell presents antigen to both the T helper and the killer but delivers co-stimulatory signals only to the helper, which is thereby stimulated to produce IL-2 for use by the nearby killer. **b**, The 'dynamic' model in which the dendritic cell offers co-stimulatory signals to both cells. It initially stimulates the T helper (left), which, in turn, stimulates and 'conditions' the dendritic cell to differentiate to a state (right) where it can now directly co-stimulate the killer.

† Present address: Unità di Immunologia, I Clinica Medica, Policlinico Umberto I, Viale del Policlinico 155, 00161 Roma, Italy.

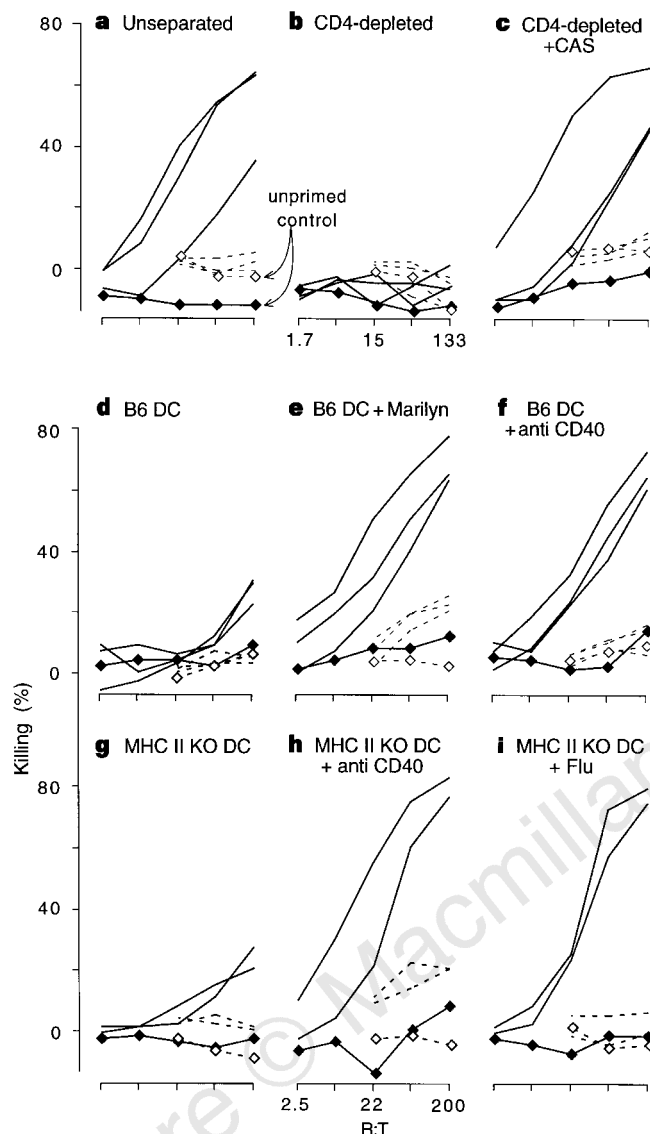


Figure 2 Four ways to help a killer. **a-c**, Role of helper cells and helper factors. Unseparated (**a**) or CD4-depleted (**b, c**) spleen cells from immunized female B6 mice were stimulated *in vitro* with B6 male spleen cells with (**c**) or without (**b**) CAS. **d-i**, Modulation of dendritic cells. CD4-depleted spleen cells from immunized female B6 mice were stimulated *in vitro* with male B6 dendritic cells (DC) (**d**), DC + Marilyn (**e**), CD40-modulated DC (**f**), unconditioned DC from MHC II KO mice (**g**), CD40-modulated DC from MHC II KO mice (**h**), or MHC II KO DC infected with influenza (Flu) (**i**). Solid lines, killing of male targets; dashed lines, killing of female targets; diamonds, unprimed controls. R:T, responder-to-target ratio.

Many other antibodies to dendritic cells did not have this effect, indicating that the CD40 molecule, rather than Fc receptors or other nonspecific changes, was responsible (not shown). We also found that conditioned dendritic cells can stimulate CD8⁺ killers from a RAG (recombination-activating gene) KO, H-Y-specific, TCR (T-cell antigen receptor) transgenic mouse that cannot generate any other T-cell subsets (not shown), thus excluding the possibility that the dendritic cells function simply by eliciting help from unconventional T-helper cells, such as the newly described NK1 subset¹⁵.

Some viruses can elicit helper-independent killer responses in CD4-deficient mice^{16,17}, whereas others, such as influenza, induce diminished but still potent responses^{18,19}. We reasoned that infected dendritic cells might undergo a change similar to that induced by T-helper cells⁶, and found that influenza-infected dendritic cells from male MHC II KO mice were indeed excellent stimulators of anti-H-Y killer cells (Fig. 2i).

Figure 3 summarizes 117 tests, showing the range and variation of responses in normal, CD4-depleted, and reconstituted cultures. Memory killers were stimulated by B6 or MHC II KO dendritic cells that were conditioned by CD4⁺ helpers, anti-CD40 modulation, or virus infection. Dendritic cells that were cultured overnight with Marilyn and then sorted by fluorescence-activated cell sorting (FACS) to remove the helper cells were as stimulatory as those in which the helpers remained. Thus a helper T cell need not linger to communicate directly with the responding killers. It can stimulate a dendritic cell and leave. What might be the relevant change in the dendritic cell?

In many cell types, modulation of CD40 induces upregulation of the B7 co-stimulatory molecules^{13,20,21}. However, our dendritic cells

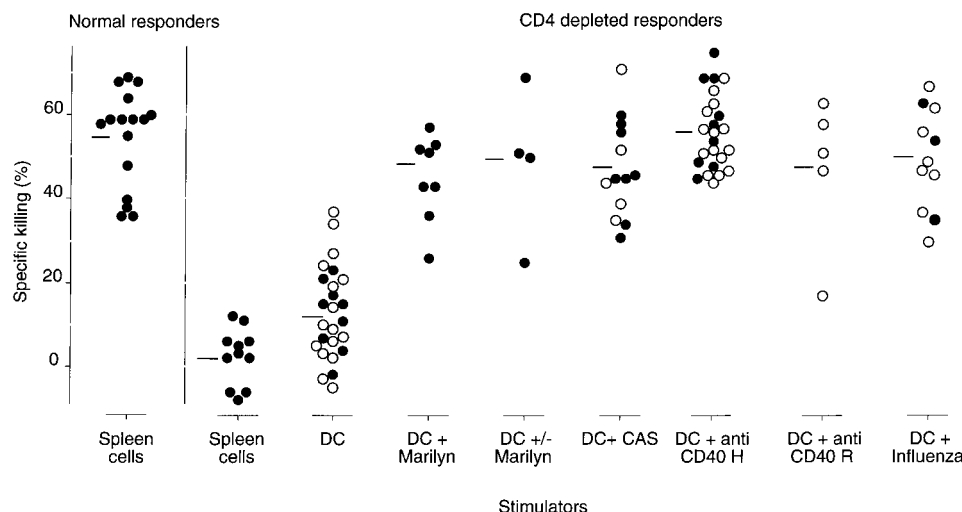


Figure 3 Summary of 117 tests showing five ways to help a killer. Spleen cells from anti-H-Y-primed female B6 mice were or were not depleted of CD4⁺ T cells and were then stimulated with spleen or dendritic cells from normal (filled circles) or MHC II KO (open circles) mice. Reading left to right: spleen cells, dendritic cells (DC), DC incubated overnight with Marilyn (+ Marilyn), DC incubated overnight with Marilyn and then sorted to remove the T cells (+/- Marilyn), DC plus 10% CAS,

DC modulated with a hamster (H) or a rat (R) anti-CD40 monoclonal antibody, DC infected with influenza virus. Each point represents the killing from a single culture at an R:T ratio at which killing shown by cultures from control mice drops off the plateau. Background killing of female targets is subtracted. Horizontal lines indicate the group average.

express very high levels of both B7.1 and B7.2 and these levels do not change with CD40 conditioning. Nevertheless, the recombinant mouse protein CTLA-4-Ig (which binds both B7 molecules) blocks their ability to stimulate killer cells, both when there is help available to the killers (Fig. 4a) and when the killers are stimulated directly (Fig. 4b). The control antibody, Ly5.2, does not block stimulation, and antibodies against B7.1 or B7.2 block only when used together, suggesting that T-killer cells may be able to use either of the two co-stimulatory molecules interchangeably. Thus, though B7 molecules are not sufficient for the stimulation of killers, they are nevertheless necessary.

Although CD40-modulated dendritic cells produce IL-6 and IL-12, these cytokines did not substitute for T-cell help (not shown). It is most likely that conditioned dendritic cells co-stimulate CD8⁺

killers directly, as they do helpers, and that this results in the production of IL-2 by the killer cell itself²². In support of this, we found that CD4-depleted memory cells were able to produce a small amount of IL-2 when stimulated by conditioned dendritic cells from MHC II KO mice (data not shown). This would explain why we can bypass help in two ways: by supplying the necessary IL-2 or by empowering the killer cell to make its own.

To determine whether conditioned dendritic cells can prime naive killers in the absence of T-cell help, we used the dendritic cells to prime killers in MHC II KO mice, which respond to some viruses^{17,18} but which had not been previously tested with helper-dependent antigens. These mice did not reject syngeneic male skin grafts (not shown) nor did they respond to injections of male spleen or dendritic cells (Fig. 5), but they did respond when primed by two injections of conditioned dendritic cells. Thus both naive and memory H-Y-specific CD8⁺ killers need help and the help can be delivered by a dendritic cell that has been conditioned by interaction with a helper, or by a virus infection.

Although there are other bodily systems in which two cells communicate through a third, most of the interactions are either mediated by soluble factors that traverse the necessary distances (such as the hypothalamic-pituitary-adrenal axis) or facilitated by stable linkages between cells (in the nervous system). The cells of the immune system use the fourth dimension, time, to transform a nearly impossible interaction between three rare, migratory and transient cells into a sequential pair of manageable two-cell engagements. This solves several problems at once. Helpers need not wastefully secrete IL-2 into their surroundings and killers need not wait for help after binding to antigen, because APCs present antigen and co-stimulatory signals simultaneously to killers, as they also do to helpers. There is also the bonus that the activity of a few helpers can be amplified, as a single T helper can 'arm' several APCs which can then, in their turn, activate a multitude of killers.

These results explain some of the contrasting findings from studies of help for killer cells. As infection by a virus can directly empower a dendritic cell to stimulate killers, some antiviral responses will depend more heavily on T helpers than others. For example, the MHC II KO mice respond normally to Sendai virus but less well to influenza. For these viruses, two types of dendritic cell may present viral antigens to CD8⁺ killers. Some dendritic cells are themselves infected and consequently directly empowered to activate killers. Others may instead be stimulated, by danger signals at the infected site²³, to pick up viral antigens. As they are not actually infected, they need helpers to become empowered and thus, in a MHC II KO mouse, will be unable to stimulate CD8⁺ killers. A prediction from this reasoning is that the helper independence of a killer response to a virus should correlate directly with that virus's ability to infect and condition the APCs.

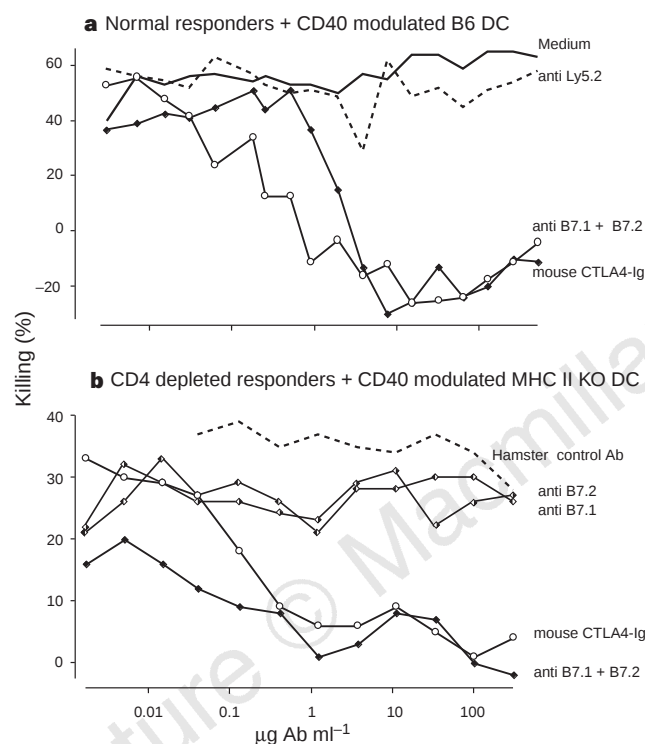


Figure 4 B7.1 and B7.2 are involved in stimulation of killer cells by conditioned dendritic cells. **a**, Unseparated, or **b**, CD4-depleted responding cells were stimulated with CD40-modulated dendritic cells (DC) from **(a)** B6 mice or **(b)** MHC II KO mice, in the presence of titrated amounts of various blocking reagents. Ab, antibody.

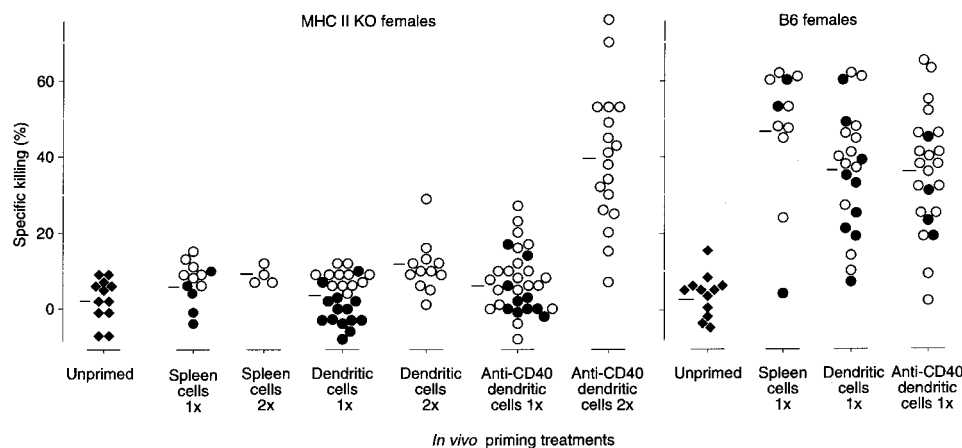


Figure 5 Virgin killers can be primed *in vivo* by CD40-modulated dendritic cells. 173 B6 or MHC II KO female mice were left untreated (diamonds) or injected once or twice with spleen or dendritic cells from B6 (filled circles) or MHC II KO (open circles) male mice. The dendritic cells were either untreated or modulated with a hamster anti-CD40 monoclonal antibody. The *in vitro* cultures contained 10% CAS, which substitutes for help and allows us to determine whether the mouse killer cells were primed *in vivo*. All mice generated killer cells against third-party CBA/J targets. Representation of killing activity is as in Fig. 3.

Conditioning by viruses may also affect responses to non-viral antigens. In an earlier study, we found helper-independent killers against the antigen Qa-1 during a hepatitis virus infection of the mouse colony⁶. For a few months of the current study, CD4-depleted mice generated moderate responses to H-Y that disappeared when we switched source colonies (though the first colony had no known pathogens). Thus, depending on the health status of a mouse, various responses may appear to be dependent or independent of T-cell help.

Finally, the existence of two different functional states of dendritic cells opens the possibility that there may be more, in analogy to the multiplicity of activation states seen with other cells of the immune system. For example, signals from functionally different T-helper cells can persuade B cells to switch to the production of different classes of antibody: T_H1 cells induce production of IgG2a; T_H2 cells elicit production of IgE and IgG1; and T_H3 cells signal the production of IgA^{24–26}. Dendritic cells may be similarly conditioned, by the signals they receive, to enter different operational states, each one initiating a different class of response. We showed here that T_H1 cells or a virus infection can empower dendritic cells to activate CD8⁺ killers. There is also evidence that IL-10 (ref. 27) or fluid from the eye²⁸ can condition APCs to become inducers of T_H2 rather than T_H1 responses. For us, the dendritic cell is beginning to look like a cell that responds to its environment in several ways and, in turn, influences several aspects of an immune response. First, it is activated by endogenous²³ or exogenous²⁹ danger signals to capture, process, and present antigen along with co-stimulatory signals and thus initiate an immune response. It is also influenced by the cells, cytokines and other signals in its environment to modify that response so that it is appropriate for both the pathogen that it is directed against and the location in which it unfolds. □

Methods

Mice and immunizations. C57BL/6 (B6) mice were purchased from Taconic farms, NY. MHC class II-knockout (MHC II KO) mice, backcrossed to B6 (N13) or C57BL/10 (N11) mice, were from the NIAID breeding contract at Taconic. Some mice were primed by an intraperitoneal injection of 3×10^6 male spleen cells or 5×10^5 dendritic cells in 200 μ l of sterile phosphate-buffered saline (PBS). Two weeks later, some of these received a second, similar injection.

Cells. For CD4 purification, spleen cells were depleted using a Midi MACs (Miltenyi, Germany) and an anti-mouse-CD4 monoclonal antibody, GK1.5, yielding less than 0.2% remaining CD4 cells by FACS analysis with the non-competing anti-CD4 monoclonal antibody RM4-4 (Pharmingen, CA).

Dendritic cells were isolated as described¹⁰.

For CD40 crosslinking, dendritic cells were incubated on ice for 10 min in PBS plus 10% mouse serum, for 20 mins with hamster (HM40-3, 5μ g ml⁻¹) or IgG2a rat (3/23, 3.5μ g ml⁻¹; Pharmingen) anti-mouse-CD40 monoclonal antibodies, and then overnight at 37 °C with goat anti-hamster or goat anti-rat antibodies (Caltag, CA) in Iscove's medium plus 10% fetal calf serum (IF10) plus 2 ng ml⁻¹ GM-CSF and 200 units ml⁻¹ IL-4. The cells were washed between and after the incubations, then injected intraperitoneally or irradiated (1,500 Rads) and used *in vitro*.

For stimulation with Marilyn, 1×10^6 male B6 dendritic cells were incubated overnight with 1.5×10^5 Marilyn, a CD4⁺ T_H1 clone specific for H-Y/A^b that was isolated from a B6 $\times$ CBA/N female mouse. These dendritic cells were irradiated and used as *in vitro* stimulators. In some experiments we removed the Marilyn, by FACS sorting with anti-H-2^k, before using the dendritic cells as stimulators. We tested for the efficiency of depletion by staining for CD4, Thy1 and T-cell antigen receptor (TCR) and by culturing an aliquot of the (unirradiated) sorted dendritic cell populations and testing the cells for proliferation. There was no evidence of contaminating Marilyn cells. For example, in two experiments the average of duplicate counts were as follows: male dendritic cells alone, experiment 1: 1,334 c.p.m., experiment 2: 1,853 c.p.m.; dendritic cells plus Marilyn, experiment 1: 25,335 c.p.m., experiment 2: 53,457 c.p.m.; dendritic cells with and then depleted of Marilyn, experiment 1: 1,114 c.p.m., experiment 2: 587 c.p.m.

For infection with influenza, dendritic cells were infected with influenza virus A/PR/8 as described³⁰, then irradiated and used as *in vitro* stimulators.

In vitro cultures. For standard cultures, 2 weeks to 1 year after *in vivo* immunization, 4×10^6 untreated or CD4-depleted spleen cells were restimulated in 2-ml cultures with 2×10^6 irradiated male spleen cells or 1.5×10^5 dendritic cells, with or without an exogenous source of mouse IL-2 (10% rat Con A supernatant, depleted of Con A, Collaborative Biomedical Products, Bedford, MA). Six days later, T-cell killing of male and female targets was tested using the JAM Test¹⁰.

For antibody-blocking cultures, anti-mouse B7.1 (17A10, hamster), anti-mouse B7.2 (2D10, rat IgG2b), CTLA-4-Ig (recombinant mouse CTLA-4-Ig fusion protein), and control antibodies anti-mouse Iy5.2 (A20.1.7, rat IgG2b, Fig. 3a) and anti-mouse MHC II (M5/114, rat IgG2b; Fig. 3b) were titrated by 1:2 in IF10, starting at 1 mg or 600 μ g ml⁻¹ in 100- μ l volumes in 96-well round-bottomed plates. 50 μ l containing 1×10^5 primed unseparated or CD4-depleted B6 female spleen cells plus 50 μ l containing 5×10^4 conditioned B6 or MHC II KO male dendritic cells were added to give a final volume of 200 μ l. Seven days later, the medium was removed and replaced with 200 μ l medium containing 10^4 target cells for the JAM Test. We also used hamster antibody UC3 and human recombinant CTLA-4-Ig. They did not block and are not reported here for clarity.

Received 13 November 1997; accepted 7 April 1998.

- Keene, J. A. & Forman, J. Helper activity is required for the *in vivo* generation of cytotoxic T lymphocytes. *J. Exp. Med.* **155**, 768–782 (1982).
- Mitchison, N. A. & O'Malley, C. Three-cell-type clusters of T cells with antigen-presenting cells best explain the epitope linkage and noncognate requirements of the *in vivo* cytolytic response. *Eur. J. Immunol.* **17**, 1579–1583 (1987).
- Bennett, S. R., Carbone, F. R., Karamalis, F., Miller, J. F. & Heath, W. R. Induction of a CD8 cytotoxic T lymphocyte response by cross-priming requires cognate CD4 T cell help. *J. Exp. Med.* **186**, 65–70 (1997).
- Guerder, S. & Matzinger, P. Activation versus tolerance: a decision made by helper T cells. *Cold Spring Harb. Symp. Quant. Biol.* **54**, 799 (1989).
- Rees, M. A., Rosenberg, A. S., Munitz, T. I. & Singer, A. *In vivo* xtal induction of antigen-specific transplantation tolerance to Qa1^b by exposure to alloantigen in the absence of T-cell help. *Proc. Natl Acad. Sci. USA* **87**, 2765–2769 (1990).
- Guerder, S. & Matzinger, P. A fail-safe mechanism for maintaining self-tolerance. *J. Exp. Med.* **176**, 553–564 (1992).
- Simpson, E. & Gordon, R. D. Responsiveness to H-Y antigen: Ir gene complementation and target cell specificity. *Immunol. Rev.* **35**, 59–75 (1977).
- Gray, D. & Matzinger, P. T cell memory is short-lived in the absence of antigen. *J. Exp. Med.* **174**, 969–974 (1991).
- Fuchs, E. J. & Matzinger, P. B cells turn off virgin but not memory T cells. *Science* **258**, 1156–1159 (1992).
- Ridge, J. P., Fuchs, E. J. & Matzinger, P. Neonatal tolerance revisited: turning on newborn T cells with dendritic cells [see comments]. *Science* **271**, 1723–1726 (1996).
- Boog, C. J. *et al.* Abolition of specific immune response defect by immunization with dendritic cells. *Nature* **318**, 59–62 (1985).
- Saeland, S., Duvert, V., Moreau, I. & Banchereau, J. Human B cell precursors proliferate and express CD23 after CD40 ligation. *J. Exp. Med.* **178**, 113–120 (1993).
- Cella, M. *et al.* Ligation of CD40 on dendritic cells triggers production of high levels of interleukin-12 and enhances T cell stimulatory capacity: T-T help via APC activation. *J. Exp. Med.* **184**, 747–752 (1996).
- Yang, Y. & Wilson, J. M. CD40 ligand-dependent T cell activation: requirement of B7-CD28 signalling through CD40. *Science* **273**, 1862–1864 (1996).
- Bendelac, A. *et al.* Mouse CD1-specific NK1 T cells: development, specificity, and function. *Annu. Rev. Immunol.* **15**, 535–562 (1997).
- Buller, R. M., Holmes, K. L., Hugin, A., Frederickson, T. N. & Morse III, H. C. Induction of cytotoxic T-cell responses *in vivo* in the absence of CD4 helper cells. *Nature* **328**, 77–79 (1987).
- Hou, S., Mo, X. Y., Hyland, L. & Doherty, P. C. Host response to Sendai virus in mice lacking class II major histocompatibility complex glycoproteins. *J. Virol.* **69**, 1429–1434 (1995).
- Tripp, R. A., Sarawar, S. R. & Doherty, P. C. Characteristics of the influenza virus-specific CD8 T cell responses in mice homozygous for disruption of the H-2A^b gene. *J. Immunol.* **155**, 2955–2959 (1995).
- Ahmed, R., Butler, L. D. & Bhatti, L. T4 T helper cell function *in vivo*: differential requirement for induction of antiviral cytotoxic T cell and antibody responses. *J. Virol.* **62**, 2102–2106 (1988).
- Yong, J. L. *et al.* Memory B cells from human tonsils colonize mucosal epithelium and directly present antigen to T cells by rapid up-regulation of B7-1 and B7-2. *Immunity* **2**, 239–248 (1995).
- Wu, Y. & Liu, Y. Viral induction of co-stimulatory activity on antigen-presenting cells bypasses the need for CD4⁺ T-cell help in CD8⁺ T-cell responses. *Curr. Biol.* **4**, 499–505 (1994).
- Paliard, X. *et al.* Simultaneous production of IL-2, IL-4, and IFN- γ by activated human CD4 and CD8 T cell clones. *J. Immunol.* **141**, 849–855 (1988).
- Matzinger, P. Tolerance, danger, and the extended family. *Annu. Rev. Immunol.* **12**, 991–1045 (1994).
- Katona, I. M., Urban, J. F. Jr, Kang, S. S., Paul, W. E. & Finkelman, F. D. IL-4 requirements for the generation of secondary *in vivo* IgE responses. *J. Immunol.* **146**, 4215–4221 (1991).
- Chen, Y., Kuchroo, V. K., Inobe, J., Hafler, D. A. & Weiner, H. L. Regulatory T cell clones induced by oral tolerance: suppression of autoimmune encephalomyelitis. *Science* **265**, 1237–1240 (1994).
- Stavnezer, J. Regulation of antibody production and class switching by TGF- β . *J. Immunol.* **155**, 1647–1651 (1995).
- Liu, L., Rich, B. E., Inobe, J., Chen, W. & Weiner, H. L. A potential pathway of Th2 development during the primary immune response: IL-10 pretreated dendritic cells prime naive CD4⁺ T cells to secrete IL-4. *Adv. Exp. Med. Biol.* **417**, 375–381 (1997).
- Wilbanks, G. A. & Streilein, J. W. Fluids from immune privileged sites endow macrophages with the capacity to induce antigen-specific immune deviation via a mechanism involving transforming growth factor- β . *Eur. J. Immunol.* **22**, 1031–1036 (1992).

29. Janeway, C. A. Jr Approaching the asymptote? Evolution and revolution in immunology. *Cold. Spring Harb. Symp. Quant. Biol.* **54**, 1–13 (1989).
30. Nonacs, R., Humborg, C., Tam, J. P. & Steinman, R. M. Mechanisms of mouse spleen dendritic cell function in the generation of influenza-specific, cytolytic T lymphocytes. *J. Exp. Med.* **176**, 519–529 (1997).

Acknowledgements. We thank C. Anderson for skin grafting; O. Lantz for the use of his unpublished TCR transgenic mouse; H. Arnheiter, A. Bendelac, L. D'Adamio, R. Germain, R. Schwartz and members of the Ghost lab (O. Alpan, C. Anderson, S. Celli, A. Frank, S. Galluci, T. Kamala, R. Weiss) for reading the manuscript and for useful suggestions; and D. Faherty, K. Hathcock and P. Linsley for blocking reagents.

Correspondence and requests for materials should be addressed to J.P.R. (e-mail: jridge@atlas.niaid.nih.gov).

Help for cytotoxic-T-cell responses is mediated by CD40 signalling

Sally R. M. Bennett^{†‡}, Francis R. Carbone[‡],
Freda Karamalis^{†‡}, Richard A. Flavell[§],
Jacques F. A. P. Miller^{*} & William R. Heath^{*}

^{*} Immunology Division, The Walter and Eliza Hall Institute of Medical Research, PO The Royal Melbourne Hospital, Melbourne, Victoria 3050, Australia

^{†‡} The Cooperative Research Centre for Vaccine Technology at the Queensland Institute of Medical Research, PO The Royal Brisbane Hospital, Herston, Queensland 4029, Australia

[‡] The Monash Medical School, Commercial Road, Prahran, Victoria 3181, Australia

[§] The Section of Immunology and Howard Hughes Medical Institute, Yale University, 310 Cedar Street, FMB 412, PO Box 208011, New Haven, Connecticut 06520-8011, USA

Cytotoxic T lymphocytes (CTLs) which carry the CD8 antigen recognize antigens that are presented on target cells by the class I major histocompatibility complex. CTLs are responsible for the killing of antigen-bearing target cells, such as virus-infected cells. Although CTL effectors can act alone when killing target cells, their differentiation from naive CD8-positive T cells is often dependent on 'help' from CD4-positive helper T (T_H) cells^{1–4}.

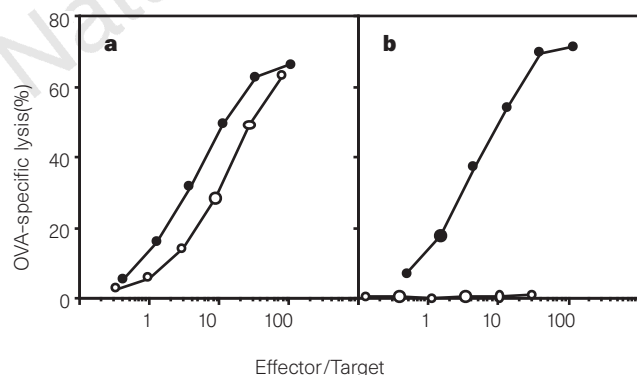


Figure 1 OVA-specific CTLs can be generated in CD4-independent and CD4-dependent ways. Normal B6 mice (filled circles) or B6 mice depleted of CD4-positive T cells by twice-weekly intraperitoneal injection of 100 μ l GK1.5 ascites⁴ (open circles) were injected either **a**, subcutaneously with 20 μ g OVAp in 200 μ l CFA, or **b**, intravenously with irradiated B6 OVA-loaded spleen cells as described⁴. After 8 days, spleen cells from each mouse were restimulated for 6 days *in vitro* as described⁴. On the day of assay, effector cells were examined for their ability to lyse ⁵¹Cr-labelled EL4 targets that were or were not pulsed with OVAp. Non-specific EL4 lysis was <10%.

Furthermore, for effective CTL priming, this help must be provided in a cognate manner, such that both the T_H cell and the CTL recognize antigen on the same antigen-presenting cell^{2,4}. One explanation for this requirement is that T_H cells are needed to convert the antigen-presenting cell into a cell that is fully competent to prime CTL⁵. Here we show that signalling through CD40 on the antigen-presenting cells can replace the requirement for T_H cells, indicating that T-cell 'help', at least for generation of CTLs by cross-priming, is mediated by signalling through CD40 on the antigen-presenting cell.

CD8-positive CTLs are responsible for the lysis of antigen-bearing target cells. These CTLs recognize peptide antigens presented by class I molecules encoded within the major histocompatibility complex (MHC). Generation of effective CTL responses often requires help from a second subset of T lymphocytes, the CD4-positive helper T (T_H) cells^{1–4}, but this is not always the case^{6,7}. In response to the soluble protein ovalbumin (OVA), both T_H -cell-dependent and T_H -cell-independent CTL immunity can be induced^{4,8}. When the K^b-restricted OVA peptide determinant spanning residues λ 57 to 264 (OVAp) was emulsified in complete Freund's adjuvant (CFA) and injected subcutaneously, CTLs could be generated in mice lacking CD4-positive T cells (Fig. 1a)⁴. In contrast, priming by intravenous injection of irradiated spleen cells loaded with OVA by osmotic shock (OVA-loaded spleen cells) required the presence of CD4-positive T cells (Fig. 1b). This latter form of immunization occurs by cross-priming^{4,9}, requiring representation of antigen by host bone-marrow-derived antigen-presenting cells (APCs). Dissection of the cellular interactions involved in this response⁴ revealed that, like the induction of CTLs that are specific for the Qa1 antigen¹, the T_H and CTL populations must recognize OVA on the same APC for effective CTL priming. This could be explained in two ways: either the T_H cells need to closely associate with the CTL to deliver short-range signals such as interleukin(IL)-2 (ref. 1), or they are required to modify the APC, converting it into a stimulatory cell for CTL priming⁵.

There is evidence that CD40 and CD40 ligand (CD154) are important in both the humoral and cellular immune responses (reviewed in refs 10, 11). These molecules have been implicated in the generation of CTL responses to adenovirus¹² and in the estab-

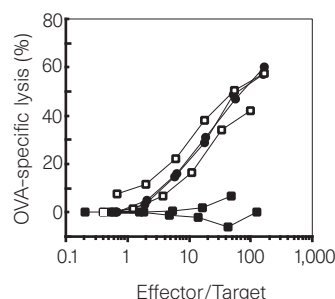


Figure 2 Treatment with a monoclonal antibody against CD40 replaces CD4-positive T-cell help in generating OVA-specific CTLs. Normal B6 mice (circles) and H-2A^b-deficient B6 mice (squares), were injected intravenously with irradiated, OVA-loaded bm1 spleen cells⁴. They were then left untreated (circles), or were injected intravenously daily for 4 days with either 0.1 mg of a CD40-specific monoclonal antibody, FGK45 (ref. 17) (open squares) or an IgG2a isotype control, KT50, specific for V α 8 (filled squares). Eight days after priming, the spleen cells from each mouse were restimulated *in vitro* for six days. We then examined their ability to lyse ⁵¹Cr-labelled EL4 targets that were or were not pulsed with OVAp. Non-specific EL4 lysis was <7%.

lishment of CTL memory to lymphocytic choriomeningitis virus¹³. CD154-deficient mice are unresponsive to adenovirus in that they do not generate either antibody or CTLs¹⁴, unless co-injected with a CD40-stimulating monoclonal antibody. However, it was not determined which cellular population was normally responsible for supplying CD154 for CTL induction; CD4-positive T_H cells, CD8-positive CTLs, or some other cell type. We reasoned, however, that as this ligand is mainly expressed by activated T_H cells^{15,16}, CD154 signalling to CD40 may represent the 'help' provided by these cells during CTL generation. To test this, we determined whether a CD40 stimulus could substitute for T_H cells in the CTL response to OVA-loaded spleen cells. Mice lacking the MHC locus H-2A^b, which do not express MHC class II molecules and therefore lack class II-restricted CD4-positive T_H cells, were primed with OVA-loaded spleen cells in the presence of either a CD40-stimulating monoclonal antibody, FGK45 (ref. 17), or an isotype-control antibody. We then examined OVA-specific CTL generation (Fig. 2). We used spleen cells from bm1 mice as the OVA-loaded priming source to ensure that stimulation could only occur through cross-priming on host APCs, as the H-2^{bm1} haplotype cannot directly present OVA to CD8-positive T cells. In this case, strong CTL responses were induced in the presence, but not the absence, of the CD40-specific monoclonal antibody. There was no CTL priming when mice were given FGK45 in the absence of OVA-loaded spleen cells (data not shown).

To exclude the possibility that the few unusual CD4-positive T cells in H-2A^b-deficient mice could substitute for T_H cells when a CD40 stimulus was provided, we performed similar experiments in thymectomized B6 mice that were depleted of CD4-positive T cells by administration of monoclonal antibody (Fig. 3). This showed that, even when CD4-positive T cells were depleted by antibody, mice could be primed only in the presence of a CD40-stimulating monoclonal antibody. These data indicate that a CD40 stimulus can substitute for T_H cells, suggesting that help is normally mediated through CD40 signalling following engagement by CD154 expressed on T_H cells.

An alternative explanation, however, is that signalling by the CD40-specific monoclonal antibody substituted for a completely different signal supplied by T_H cells. To exclude this possibility, we examined OVA-specific CTL responses in CD40-deficient or CD154-deficient mice. If CD40 signalling is normally involved in this form of priming, these mice should not generate OVA-specific

CTL responses. CTL responses were indeed poor in both CD154-deficient mice (Fig. 4a) and CD40-deficient mice (Fig. 4b). In one of four experiments, using CD40-deficient mice, we detected relatively efficient CTL responses (data not shown). As we performed this experiment soon after introducing the CD40-deficient mice into our colony, we suspect that subclinical infections provided inflammatory signals that substituted for the CD154 signal of T_H cells, in much the same way as CFA replaced the need for T_H cells in the generation of OVA-specific CTLs^{4,8} (Fig. 1a). The general lack of CTL responses in mice lacking either CD40 or CD154 support a role for CD40 signalling in the cross-priming of OVA-specific CTLs.

Priming of CTLs by OVA-loaded spleen cells occurs through antigen presentation by host APCs. This priming requires help mediated by a T_H-cell interaction with the same APCs as seen by the CTLs⁴. The requirement for CD40 and CD154 in this response (Fig. 4), and the ability of a CD40-specific monoclonal antibody to replace T_H cells (Figs 2, 3), indicates that CD4-positive T-cell help mainly supplies CD154 for CD40 signalling. As CD40 is expressed by APCs such as dendritic cells (reviewed in ref. 10), and as signalling through CD40 is important for maturation of several antigen-presenting functions^{14,18–25}, the cross-priming APC probably represents the target cell that is stimulated through CD40. B cells, which also express CD40, are unlikely to be involved in this response, as presentation of antigen by this population is tolerogenic for naive CD8-positive T cells, even in the presence of T_H cells²⁶. A small proportion of CD8-positive T cells has been reported to express CD40 (ref. 27), so a direct interaction between CD154 on the T_H cell and CD40 on the CTL may occur. This would, however, require interaction between these cells while associated on the APC, as murine CTLs do not express MHC class II molecules.

How signalling through CD40 enables the cross-priming APC effectively to stimulate CTLs is unclear, although CD40 signalling may lead to the upregulation of co-stimulatory molecules such as B7 (refs 14, 19), CD44H (ref. 20) and ICAM-1 (ref. 21) on various cell types, and to the secretion of cytokines, such as tumour-necrosis factor- α , IL-1 β and IL-12 (refs 10, 19, 22–25). Whether these or other events form the critical downstream helper effects provided by T_H cells is unknown. Although CD40–CD154 interactions are important for the induction of CTL immunity to adenovirus¹², they are not essential for induction of CTL responses to several other viruses^{13,28}. We suggest that, like CFA, which allows the induction of T_H-cell-independent CTL responses to OVA, some viruses may

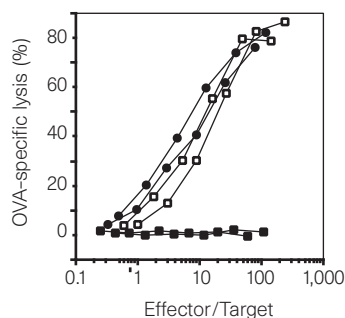


Figure 3 A CD40-specific monoclonal antibody can replace CD4-positive T-cell help in the induction of OVA-specific CTLs. Thymectomized B6 mice were depleted of CD4-positive T cells by twice-weekly intraperitoneal injection of GK1.5 ascites⁴ (squares) or were left untreated (circles). All mice were then immunized with 25×10^6 OVA-loaded B6 spleen cells. CD4-depleted mice were then either treated with 0.1 mg anti-CD40 antibody intravenously daily for 7 days (open squares), or left untreated (filled squares). Spleen cells were then cultured *in vitro* for 6 days with 10^7 irradiated E.G7 cells. The resulting effector cells were then included in a chromium-release assay using EL4 target cells that were or were not pulsed with OVAp. Nonspecific EL4 lysis was <5%.

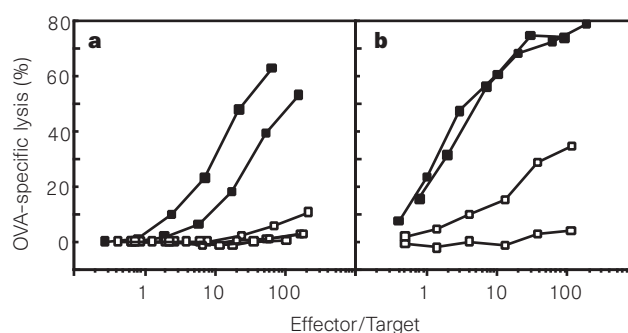


Figure 4 The role of CD154 and CD40 in the generation of OVA-specific CTLs by cross-priming. **a**, We primed four CD154-deficient mice²⁹ (open squares) and two normal B6 mice (filled squares) with bm1 OVA-loaded spleen cells, and then examined for CTL generation. We performed this experiment twice, with similar results. **b**, We immunized two CD40-deficient³⁰ mice (open squares) and two normal B6 mice (filled squares) intravenously with bm1 OVA-loaded spleen cells, and then examined for CTL generation. We performed this experiment four times, with three experiments giving similar results. Nonspecific EL4 lysis was <36%.

provide inflammatory signals that activate APCs directly, circumventing their need for activation through CD40. Perhaps the requirement for CD40 signalling depends on whether the professional APC is directly infected or presents antigen indirectly by cross-presentation. Alternatively, there may be other inherent properties of some viruses that allow CD40-independent activation of APCs.

The ability to replace T_H cells with an antibody against CD40 may provide an opportunity for vaccination under circumstances that were previously difficult, either because of a lack of available class II-restricted antigenic determinants (as may occur with some tumour antigens) or because of a deficiency in CD4-positive T_H cells (as occurs in AIDS patients). Further dissection of the molecular interactions involved in the generation of immunogenic CTL responses may help in vaccine design and function. □

Methods

Generation of CTL responses. CTL responses to OVA were generated as described⁴. Briefly, mice were primed with 25×10^6 OVA-loaded spleen cells. Seven to eight days later, we restimulated spleen cells from primed mice *in vitro* for six days with irradiated (1,500 centiGray) OVA-loaded B6 spleen cells or irradiated (20,000 centiGray) E.G7 tumour cells. We then examined the ability of the primed spleen cells to lyse ⁵¹Cr-labelled EL4 target cells, pulsed or unpulsed with OVA as described⁴. The percentage of OVA-specific lysis was then calculated as the percentage of OVA-pulsed EL4 lysis minus the percentage of EL4 lysis. To generate CTLs in CD154-deficient²⁹ or CD40-deficient³⁰ mice, we modified the *in vitro* culture conditions slightly to avoid responses to mouse strain 129 minor histocompatibility antigens. Seven days after immunization with OVA-loaded bm1 spleen cells, we removed 'responder' spleens and irradiated half the cells from each spleen (1,500 centiGray). We then loaded these irradiated cells with OVA with OVA by osmotic shock and used these cells as *in vitro* stimulators for the remaining splenocytes. We combined unirradiated 'responder' spleen cells (3×10^6 cells ml⁻¹) with irradiated OVA-loaded cells from the same animal (3×10^6 cells ml⁻¹) and cultured them *in vitro* for six days in several 2-ml cultures. On day 6, we pooled effector cells from each mouse and used them in a chromium-release assay as described⁴.

Received 20 November 1997; accepted 16 March 1998.

- Keene, J. A. & Forman, J. Helper activity is required for the *in vivo* generation of cytotoxic T lymphocytes. *J. Exp. Med.* **155**, 768–782 (1982).
- von Herrath, M. G., Yokoyama, M., Dockter, J., Oldstone, M. B. & Whitton, J. L. CD4-deficient mice have reduced levels of memory cytotoxic T lymphocytes after immunization and show diminished resistance to subsequent virus challenge. *J. Virol.* **70**, 1072–1079 (1996).
- Cardin, R. D., Brooks, J. W., Sarawar, S. R. & Doherty, P. C. Progressive loss of CD8⁺ T cell-mediated control of a γ -herpesvirus in the absence of CD4⁺ T cells. *J. Exp. Med.* **184**, 863–871 (1996).
- Bennett, S. R., Carbone, F. R., Karamalis, F., Miller, J. E. A. P. & Heath, W. R. Induction of a CD8 cytotoxic T lymphocyte response by cross-priming requires cognate CD4 help. *J. Exp. Med.* **186**, 65–70 (1997).
- Guerder, S. & Matzinger, P. A fail-safe mechanism for maintaining self-tolerance. *J. Exp. Med.* **176**, 553–564 (1992).
- Butler, R. M., Holmes, K. L., Hugin, A., Frederickson, T. N. & Morse, H. C. Induction of cytotoxic T-cell responses *in vivo* in the absence of CD4 helper cells. *Nature* **328**, 77–79 (1987).
- Ahmed, R., Butler, L. D. & Bhatti, L. T4⁺ T helper cell function *in vivo*: differential requirement for induction of antiviral cytotoxic T-cell and antibody responses. *J. Virol.* **62**, 2102–2106 (1988).
- Ke, Y. & Kapp, J. A. Oral antigen inhibits priming of D8⁺ CTL, CD4⁺ T cells, and antibody responses while activating D8⁺ suppressor T cells. *J. Immunol.* **156**, 916–921 (1996).
- Carbone, F. R. & Bevan, M. J. Class I-restricted processing and presentation of exogenous cell-associated antigen *in vivo*. *J. Exp. Med.* **171**, 377–387 (1990).
- Stout, R. D. & Suttles, J. The many roles of CD40 in cell-mediated inflammatory responses. *Immunol. Today* **17**, 487–492 (1996).
- Grewal, I. S. & Flavell, R. A. A central role of CD40 ligand in the regulation of CD4⁺ T-cell responses. *Immunol. Today* **17**, 410–414 (1996).
- Yang, Y. et al. Transient subversion of CD40 ligand function diminishes immune responses to adenovirus vectors in mouse liver and lung tissues. *J. Virol.* **70**, 6370–6377 (1996).
- Borrow, P. et al. CD40L-deficient mice show deficits in antiviral immunity and have an impaired memory CD8⁺ CTL response. *J. Exp. Med.* **183**, 2129–2142 (1996).
- Yang, Y. & Wilson, J. M. CD40 ligand-dependent T cell activation: requirement of B7-CD28 signaling through CD40. *Science* **273**, 1862–1864 (1996).
- Roy, M., Waldschmidt, T., Aruffo, A., Ledbetter, J. A. & Noelle, R. J. The regulation of the expression of gp39, the CD40 ligand, on normal and cloned CD4⁺ T cells. *J. Immunol.* **151**, 2497–2510 (1993).
- Sad, S. et al. Cytotoxicity and weak CD40 ligand expression of D8⁺ type 2 cytotoxic T cells restricts their potential B cell helper activity. *Eur. J. Immunol.* **27**, 914–922 (1997).
- Rolink, A., Melchers, F. & Andersson, J. The SCID but not the RAG-2 gene product is required for S mu-5 epsilon heavy chain class switching. *Immunology* **5**, 319–330 (1996).
- Sallusto, F., Cella, M., Danieli, C. & Lanzavecchia, A. Dendritic cells use macropinocytosis and the mannose receptor to concentrate macromolecules in the major histocompatibility complex class II compartment: downregulation by cytokines and bacterial products. *J. Exp. Med.* **182**, 389–400 (1995).
- Grewal, I. S. et al. Requirement for CD40 ligand in costimulation induction, T cell activation, and experimental allergic encephalomyelitis. *Science* **273**, 1864–1867 (1996).

- Guo, Y. et al. Identification of a costimulatory molecule rapidly induced by CD40L as CD44H. *J. Exp. Med.* **184**, 955–961 (1996).
- Shinde, S. et al. CD40L is important for induction of, but not response to, costimulatory activity. ICAM-1 as the second costimulatory molecule rapidly up-regulated by CD40L. *J. Immunol.* **157**, 2764–2768 (1996).
- Alderson, M. R. et al. CD40 expression by human monocytes: regulation by cytokines and activation of monocytes by the ligand for CD40. *J. Exp. Med.* **178**, 669–674 (1993).
- Wagner, D. L., Stout, R. D. & Suttles, J. Role of the CD40-CD40 ligand interaction in CD4⁺ T cell contact-dependent activation of monocyte interleukin-1 synthesis. *Eur. J. Immunol.* **24**, 3148–3154 (1994).
- Shu, U. et al. Activated T cells induce interleukin-12 production by monocytes via CD40-CD40 ligand interaction. *Eur. J. Immunol.* **25**, 1125–1128 (1995).
- Kiener, P. A. et al. Stimulation of CD40 with purified soluble gp39 induces proinflammatory responses in human monocytes. *J. Immunol.* **155**, 4917–4925 (1995).
- Fuchs, E. J. & Matzinger, P. B cells turn off virgin but not memory T cells. *Science* **258**, 1156–1159 (1992).
- Armitage, R. J. et al. CD40 ligand is a T cell growth factor. *Eur. J. Immunol.* **23**, 2326–2331 (1993).
- Whitmire, J. K., Slifka, M. K., Grewal, I. S., Flavell, R. A. & Ahmed, R. CD40 ligand-deficient mice generate a normal primary cytotoxic T-lymphocyte response but a defective humoral response to a viral infection. *J. Virol.* **70**, 8375–8381 (1996).
- Xu, J. et al. Mice deficient for the D40 ligand. *Immunity* **1**, 423–431 (1994).
- Kawabe, T. et al. The immune responses in D40-deficient mice: impaired immunoglobulin class switching and germinal center formation. *Immunity* **1**, 167–178 (1994).

Acknowledgements. We thank H. Kikutani for the use of CD40-deficient mice; J. Ruby for providing these mice; A. Rolink for provision of the CD40-specific monoclonal antibody FGK145; D. Mathis for the use of class II deficient mice; J. Falso and T. Banjanin for technical assistance; and P. Matzinger, J. Ridge, and S. Schoenberger for helpful discussion. This work was supported by the Cooperative Research Centre for Vaccine Technology, and grants from the NIH, National Health and Medical Research Council, and Australian Research Council. R.A.F. is an Investigator of the Howard Hughes Medical Institute.

Correspondence and requests for materials should be addressed to W.R.H. (e-mail: Heath@wehi.edu.au) or to F.R.C. (e-mail: carbone@cobra.path.monash.edu.au).

T-cell help for cytotoxic T lymphocytes is mediated by CD40-CD40L interactions

Stephen P. Schoenberger^{*†‡}, Rene E. M. Toes^{*†}, Ellen I. H. van der Voort^{*}, Rienk Offringa^{*} & Cornelis J. M. Melief^{*}

^{*} Department of Immunohematology and Blood Bank, University Hospital Leiden, PO Box 9600, 2333 AZ, Leiden, The Netherlands

[†] These authors contributed equally to this work.

Although *in vivo* priming of CD8⁺ cytotoxic T lymphocytes (CTLs) generally requires the participation of CD4⁺ T-helper lymphocytes^{1,2}, the nature of the 'help' provided to CTLs is unknown³. One widely held view is that help for CTLs is mediated by cytokines produced by T-helper cells activated in proximity to the CTL precursor at the surface of an antigen-presenting cell (APC)⁴. An alternative theory is that, rather than being directly supplied to the CTL by the helper cell, help is delivered through activation of the APC, which can then prime the CTL directly⁵. CD40 and its ligand, CD40L, may activate the APC to allow CTL priming. CD40L is expressed on the surface of activated CD4⁺ T-helper cells and is involved in their activation and in the development of their effector functions^{6,7}. Ligation of CD40 on the surface of APCs such as dendritic cells, macrophages and B cells greatly increases their antigen-presentation and co-stimulatory capacity^{8–11}. Here we report that signalling through CD40 can replace CD4⁺ T-helper cells in priming of helper-dependent CD8⁺ CTL responses. Blockade of CD40L inhibits CTL priming; this inhibition is overcome by signalling through CD40. CD40-CD40L interactions are therefore vital in the delivery of T-cell help for CTL priming.

The dependence on 'help' mediated by CD4⁺ T cells for the priming of antigen-specific CD8⁺ CTL responses has been well documented^{1–4}. Despite this, the mechanism through which help is

[‡] Present address: Division of Immune Regulation, La Jolla Institute for Allergy and Immunology, San Diego, California 92121, USA.

provided to CTLs has remained unknown. We used a well characterized model system to probe the mechanism of T-cell help for the primary activation of CD8⁺ CTL responses *in vivo*. C57BL/6 (B6; with the major histocompatibility complex (MHC) H-2^b) mice immunized with allogeneic BALB/c (H-2^d) mouse embryo cells (MECs) expressing the human adenovirus type 5 early region 1 (Ad5E1-BALB/c MECs) generated strong CTL responses against an H-2D^b-restricted epitope of the adenovirus E1B protein (E1B₁₉₂₋₂₀₀) (Fig. 1a)¹². As the allogeneic H-2^d MHCs expressed by the immunizing MECs cannot prime H-2^b-restricted host CTLs¹², generation of E1B-specific CTLs must require cross-priming, that is, the uptake and H-2^b-restricted re-presentation of antigen by host APCs^{13,14}. Cross-priming of E1B-specific CTLs is strictly helper-dependent (Fig. 1b), as mice depleted of CD4⁺ T-helper (T_H) cells before immunization no longer mounted an E1B-specific CTL response.

Studies of the requirement for T-cell help in cross-priming of CTLs have shown that both T_H cells and CTLs must recognize antigens presented on the same APC^{3,4}. Although this could be explained by a proximity requirement for the efficient delivery from T_H cells of soluble factors such as interleukin (IL)-2, another possibility is that a cognate interaction between T_H cells and APCs is required to convert the APC to a state that is capable of directly priming naive CTLs⁵. CD40 and CD40L (CD154) may mediate the activation of APCs by T_H cells. CD40 (expressed on APCs) and CD40L (expressed on activated T_H cells) are important in both humoral and cellular immune responses (reviewed in refs 15, 16). A failure to properly activate APCs could explain many of the immune defects of CD40L-deficient mice, such as the inability to prime antigen-specific T cells, to resist infection with *Leishmania*, or to reject tumours following vaccination with tumour cells^{6,17,18}.

To investigate whether signalling through CD40 can replace CD4⁺ T_H cells in CTL priming, we depleted mice of CD4⁺ cells *in vivo* before immunization with Ad5E1-BALB/c MECs. One day after immunization, the mice received a single injection of an anti-CD40 activating antibody, FGK45 (ref. 19), or of an isotype-matched control antibody. Administration of FGK45 to CD4-depleted, immunized mice resulted in the efficient restoration of E1B-specific CTL responses (Fig. 2a) whereas treatment with the control antibody did not (Fig. 2b). Priming of E1B-specific CTLs was not detected in naive mice treated with FGK45 alone (not shown). To address the possibility that the effect of FGK45 was mediated through remaining CD4⁺ cells that were not depleted by treatment with the anti-CD4 antibody, we immunized B6 I-A^b knockout mice, which lack functional MHC class II-restricted CD4⁺ periph-

eral T cells²⁰, with the Ad5E1-BALB/c MECs. The response to immunization in these mice mirrors that seen in the CD4-depleted mice, in that E1B-specific CTLs were detectable only in mice receiving the CD40-activating antibody (Fig. 2c), and not in those receiving the control antibody (Fig. 2d). We studied whether the requirement for anti-CD40 antibodies in priming of CTLs in CD4-depleted mice could be replaced by administration of bacterial lipopolysaccharide (LPS) (50 µg intravenous), a potent inducer of proinflammatory cytokines, or IL-2 (1 × 10⁵ Cetus units in incomplete Freund's adjuvant, subcutaneous) following immunization with Ad5E1-BALB/c MECs. Whereas CD4-depleted mice treated with FGK45 exhibited strong E1B-specific CTL activity, neither LPS nor IL-2 treatment resulted in detectable CTL priming (not shown). We also observed no binding of FGK45 to CD8⁺ CTLs from either naive or immunized mice, although significant binding to surface-immunoglobulin-positive B cells was consistently observed (not shown). These results indicate that activation through CD40 can bypass the requirement for CD4⁺ T-helper cells in the cross-priming of E1B-specific CTLs.

Ligation of CD40 on B cells upregulates their costimulatory activity^{11,21}, suggesting a role for these cells in the restoration of CTL priming by treatment with anti-CD40 antibodies. To address this question, we immunized B6 µMT mice, which lack mature B cells²², with the allogeneic Ad5E1-BALB/c MECs. Cross-priming of E1B-specific CTLs did not require mature B cells (Fig. 3a). However, when depleted of CD4⁺ cells, the B-cell-deficient mice did not generate an E1B-specific CTL response (Fig. 3b). Activation through CD40 with FGK45 completely restored the capacity of the CD4-depleted µMT mice to prime E1B-specific CTLs (Fig. 3c). Thus B cells are not required as APCs or accessory cells for cross-priming in our system, nor are they required for the CD40-mediated restoration of cross-priming of CTLs in the absence of CD4⁺ T-helper cells.

Our results thus far are consistent with the idea that FGK45-mediated restoration of CTL priming operates through activation

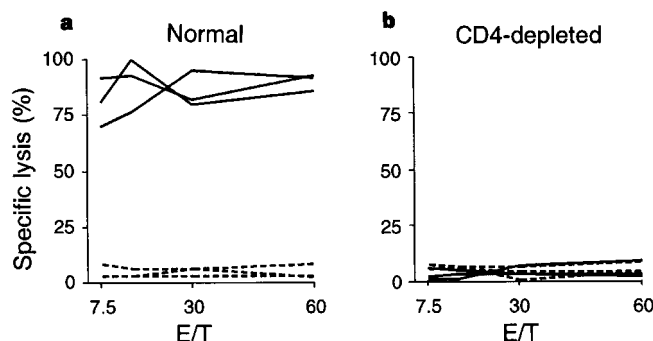


Figure 1 Cross-priming of E1B-specific CTLs requires CD4⁺ T-helper cells. Splenocytes from **a**, normal, or **b**, CD4-depleted B6 mice immunized with Ad5E1-BALB/c MECs were tested at various effector/target (E/T) ratios for lysis of syngeneic MEC target cells loaded with the E1B₁₉₂₋₂₀₀ peptide (solid lines) or a D^b-restricted control peptide, HPV-16 E7₄₉₋₅₇ (dashed lines). Each line represents one mouse. Data shown represent one experiment of three performed with similar results.

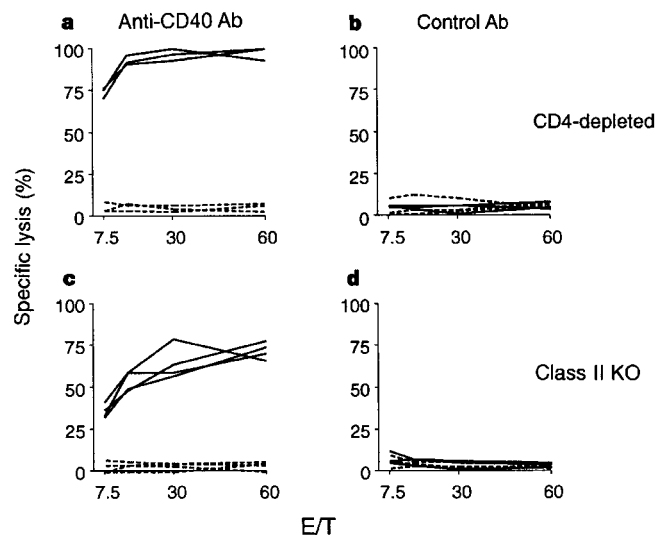


Figure 2 CD40 activation replaces CD4⁺ helper T cells. Splenocytes from **a**, **b**, CD4-depleted, or **c**, **d**, class-II-deficient I-A^b-knockout (KO) B6 mice were immunized with Ad5E1-BALB/c MECs and treated with either the CD40-activating antibody (Ab) FGK45 (**a**, **c**), or an isotype control antibody (**b**, **d**). These splenocytes were tested for E1B-specific CTL activity on syngeneic MEC target cells loaded with either the E1B₁₉₂₋₂₀₀ peptide (solid lines) or the HPV E7₄₉₋₅₇ control peptide (dashed lines). Each line represents a single mouse. Data depicted are derived from two independent experiments. E/T, effector/target ratio.

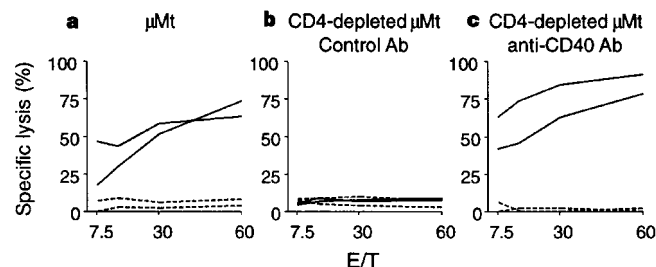


Figure 3 B cells are not essential as cross-priming APCs or for anti-CD40-mediated restoration of cross-priming. Splenocytes were taken from **a**, untreated, or **b**, **c**, CD4-depleted B-cell-deficient μ Mt B6 mice which were immunized with Ad5E1-BALB/c MECs and which received either an isotype control antibody (**b**) or the CD40-activating antibody FGK45 (**c**). These splenocytes were tested for E1B-specific CTL activity on syngeneic MEC target cells loaded with either the E1B₁₉₂₋₂₀₀ peptide (solid lines) or the HPV E7₄₉₋₅₇ control peptide (dashed lines). Each line represents one mouse. Data shown represent one experiment of two performed with similar results. E/T, effector/target ratio μ Mt, B-cell-deficient mice.

of CD40 on a non-B-cell APC, most likely of the dendritic cell/macrophage lineage. If this represents a physiological pathway used by CD4⁺ T cells to help CTLs, blocking the ability of the CD4⁺ T cells to interact with APC through CD40L-CD40 interactions would be expected to diminish priming of E1B-specific CTL responses in normal mice. We immunized B6 mice with Ad5E1-BALB/c MECs and then treated them with either a CD40L-blocking antibody, MR1 (ref. 23), or control antibodies. Blockade of CD40L results in drastically reduced E1B-specific CTL responses (Fig. 4a) compared with the efficient CTL priming seen in mice receiving the control antibodies (Fig. 4b). The priming defect induced by CD40L blockade was fully restored following CD40 signalling by FGK45 (Fig. 4c). Thus the defect induced by CD40L blockade lies in the failure of T_H cells to transmit, rather than to receive, CD40L-mediated signals.

Our results reaffirm the notion that priming of virgin CTLs *in vivo* requires the interaction of three distinct cell types: the antigen-specific CD4⁺ and CD8⁺ T cells and the antigen-bearing APC. However, in contrast to the model proposed by Keene and Forman⁴, our results indicate that, rather than directly stimulating the CTL, the contribution of CD4⁺ T_H cells involves the activation of professional (non-B-cell) APCs through CD40-CD40L interactions. These data support the model of T-cell help for CTLs that was first proposed by Guerder and Matzinger⁵ and identify CD40-CD40L interactions as mediating the activation of APC by T_H cells. This mechanism of T-cell help allows the interaction of three cell types to be accomplished through sequential, two-cell interactions, the first involving T_H-mediated activation of APCs, which are then able to prime CTLs directly.

The recovery of CTL responses by treatment with an activating antibody against CD40 has been seen by infecting CD40L-deficient mice with a recombinant adenovirus²⁴. However, in this study the role of T_H cells in CTL priming was not addressed, as the CD40L-deficient mice used were not depleted of CD4⁺ T cells. Here we have shown that anti-CD40 antibodies allow the priming of helper-dependent CTLs in the physical, or functional, absence of CD4⁺ cells.

Although the identity of the host APCs involved in cross-priming remains unknown, our results rule out a requirement for B cells in our system, either as APCs or as sources of nonspecific cytokine/costimulatory signals in response to CD40 activation. Thus, the effect of anti-CD40 antibodies may be mediated through the activation of a CD40⁺ non-B-cell APC. Among the other CD40-expressing APCs, dendritic cells are those most capable of priming T cells²⁵. It is therefore likely that the mechanism by which CD40

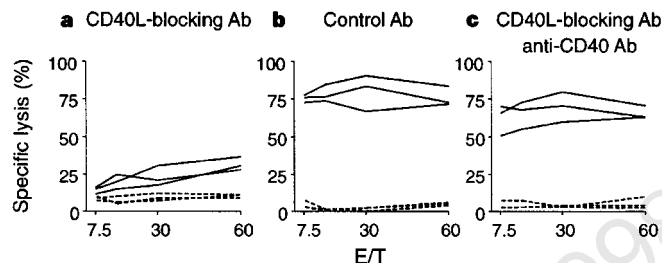


Figure 4 CD40L blockade prevents cross-priming of E1B-specific CTLs. Splenocytes were taken from **a**, **b**, B6 mice immunized with Ad5E1-BALB/c MECs and treated with **a**, the CD40L-blocking antibody MR1, or **b**, hamster IgG control antibodies, or **c**, from mice treated with the CD40L-blocking antibody that received the CD40-activating antibody FGK45 24 h after immunization. These splenocytes were tested for E1B-specific CTL activity on syngeneic MEC target cells loaded with the E1B₁₉₂₋₂₀₀ peptide (solid lines) or the HPV E7₄₉₋₅₇ control peptide (dashed lines). Each line represents one mouse. Data shown represent one experiment of two performed with similar results. E/T, effector/target ratio.

signalling allows helper-independent CTL priming involves one or more of the alterations that are induced by ligation of CD40 on dendritic cells, such as upregulation of surface CD80/CD86 levels or the production of IL-12 (refs 8, 9). Conversely, the mechanism of CD40-mediated CTL priming may involve the interruption of negative signals transmitted by the APC.

Our results provide new insights into the nature of helper-dependent CTL responses and suggest methods for their manipulation. Vaccination strategies based on the activation of APCs through CD40 may be useful in eliciting CTL responses in situations where the number or activity of CD4⁺ T_H cells is limiting, such as occurs in AIDS patients. These findings are also relevant to tumour immunology, as they indicate that a lack of tumour-specific T_H-cell determinants may prevent efficient priming of tumour-specific CTLs. By combining active vaccination with CD40 signalling *in vivo*, however, a more effective anti-tumour immune response may be realized. CD40 signalling may also be useful in achieving the full CTL-stimulatory potential of dendritic-cell-based vaccines prepared *in vitro*²⁶. Finally, our data contribute to knowledge of the emerging role of CD40-CD40L interactions as key elements in the regulation of both T- and B-cell immune responses. The linkage of CTL priming to T_H-dependent APC activation imposes an important regulatory control point in the induction of immune responses against self and non-self antigens. Cross-presentation of ovalbumin (OVA) by bone-marrow-derived APCs has been shown to lead to deletion of OVA-specific CTLs when OVA is expressed as a self antigen on healthy tissues²⁷. This contrasts with the activation of OVA-specific CTLs by cross-priming APCs, which occurs when OVA-loaded cells are used for immunization⁴. Our results indicate that the activation state of the APC, which depends on a cognate interaction with T_H cells mediated by CD40-CD40L interactions, may be important in orchestrating these different outcomes in the response to the same antigen. A new understanding of the importance of CD40-CD40L interactions for CTL priming and tolerance should allow the development of new methods for both vaccination and tolerance induction. □

Methods

Mice, cell lines, and immunizations. 6–12-week-old male B6 (H-2^b) mice used in this study were obtained from the Netherlands Cancer Institute (Amsterdam). B6I-A^b-knockout and IgM-knockout (μ MT) mice were obtained from P. Matzinger. The generation of Ad5E1-BALB/c (H-2^d) MECs and Ad5E1-B6 MECs has been described^{13,28}. All *in vitro* cultures were grown in Iscove's modified Dulbecco's medium (IMDM) (Gibco, Paisley)

supplemented with 8% fetal calf serum, 5×10^{-5} M 2-mercaptoethanol, glutamine and penicillin (culture medium). B6 mice were immunized with 1×10^7 irradiated (100 Gy) Ad5E1-BALB/c MECs in 0.2 ml PBS injected subcutaneously in the flank. Two weeks later, spleens were removed and single-cell suspensions prepared by mechanical disruption. 5×10^6 unseparated splenocytes were restimulated with 5×10^5 irradiated Ad5E1-B6 MECs in 2-ml cultures in 24-well plates. No additional growth factors or cytokines were added to these cultures.

Cytotoxicity assays. Following a 6-day culture, viable lymphocytes were isolated from splenocyte cultures and used as effectors in europium-release cytotoxicity assays as described¹². The mean percentage specific lysis of triplicate cultures was calculated as follows: specific lysis = [(c.p.m. experimental release - c.p.m. spontaneous release)/(c.p.m. maximum release - c.p.m. spontaneous release)] $\times$ 100. Spontaneous release was always <20%. Peptides were synthesized using solid-phase chemistry on a Abimed AMS 422 peptide synthesizer.

In vivo antibody treatment. Mice were depleted of CD4⁺ cells by three consecutive intraperitoneal (i.p.) injections of 100 μ g of the anti-CD4 antibody GK1.5 (ref. 29) given every other day before immunization. Depletion was confirmed by fluorescence-activated cell sorting analysis using the RM4-4 antibody (Pharmingen), which does not compete with GK1.5 for binding to CD4. Depleted populations contained <0.25% CD4 cells. Depletion of CD4 cells was maintained by administration of 100 μ g GK1.5 every 5–7 days until spleens were removed two weeks after immunization.

For CD40 activation, mice received 100 μ g purified CD40-activating antibody FGK45 or a rat IgG2a control antibody (Biosource, Belgium) given intravenously in 0.2 ml PBS 24 h after immunization with Ad5E1-BALB/c MECs.

To block CD40L, mice received 250 μ g HPLC-purified CD40L-blocking antibody MR1 (ref. 23) or hamster IgG control antibodies in 0.2 ml PBS given intraperitoneally every other day for 11 days beginning on the day of immunization with Ad5E1-BALB/c MECs.

Received 10 November 1997; accepted 7 April 1998.

1. Simpson, E. & Gordon, R. D. Responsiveness to H-Y antigen. Ir-gene complementation and target cell specificity. *Immunol. Rev.* **35**, 59–75 (1997).
2. Husmann, L. A. & Bevan, M. J. Cooperation between helper T cells and cytotoxic T lymphocyte precursors. *Ann. NY Acad. Sci.* **532**, 158–169 (1988).
3. Bennett, S. R. M., Carbone, F. R., Karamalis, F., Miller, J. F. A. P. & Heath, W. R. Induction of a CD8+ cytotoxic T lymphocyte response by cross-priming requires cognate CD4+ T cell help. *J. Exp. Med.* **186**, 65–70 (1997).
4. Keene, J. & Forman, J. Helper activity is required for the *in vivo* generation of cytotoxic T lymphocytes. *J. Exp. Med.* **155**, 768–782 (1982).
5. Guerder, S. & Matzinger, P. A fail-safe mechanism for maintaining self-tolerance. *J. Exp. Med.* **176**, 553–564 (1992).
6. Grewal, I. S., Xu, J. & Flavell, R. A. Impairment of antigen-specific T-cell priming in mice lacking CD40 ligand. *Nature* **378**, 617–620 (1995).
7. van Essen, D., Kikutani, H. & Gray, D. CD40 ligand-transduced co-stimulation of T cells in the development of helper function. *Nature* **378**, 620–623 (1995).
8. Koch, F. *et al.* High level IL-12 production by murine dendritic cells: upregulation via MHC class II and CD40 molecules and downregulation by IL-4 and IL-10. *J. Exp. Med.* **184**, 741–746 (1996).
9. Cella, M. *et al.* Ligation of CD40 on dendritic cells triggers production of high levels of interleukin-12 and enhances T cell stimulatory capacity: T-T help via APC activation. *J. Exp. Med.* **184**, 747–752 (1996).
10. Stout, R. D., Suttles, J., Xu, J.-C., Grewal, I. S. & Flavell, R. A. Impaired T cell-mediated macrophage activation in CD40 ligand-deficient mice. *J. Immunol.* **156**, 8–11 (1996).
11. Kennedy, M. K. *et al.* Induction of B cell costimulatory function by recombinant murine CD40 ligand. *Eur. J. Immunol.* **24**, 116–123 (1994).
12. Toes, R. E. M. *et al.* Protective anti-tumor immunity induced by immunization with completely allogeneic tumor cells. *Cancer Res.* **56**, 3782–3787 (1996).
13. Bevan, M. J. Cross-priming for a secondary cytotoxic response to minor H antigens with H-2 congenic cells which do not cross-react in the cytotoxic assay. *J. Exp. Med.* **143**, 1283–1288 (1976).
14. Huang, A. Y. *et al.* Role of bone marrow-derived cells in presenting MHC class I-restricted tumor antigens. *Science* **264**, 961–965 (1994).
15. Grewal, I. S. & Flavell, R. A. A central role of CD40 ligand in the regulation of CD4⁺ T-cell responses. *Immunol. Today* **17**, 410–414 (1996).
16. Noelle, R. CD40 and its ligand in host defense. *Immunity* **4**, 415–419 (1996).
17. Soong, L. *et al.* Disruption of CD40–CD40 ligand interactions results in an enhanced susceptibility to *Leishmania amazonensis* infection. *Immunity* **4**, 263–273 (1996).
18. Mackey, M. F. *et al.* Protective immunity induced by tumor vaccines requires interactions between CD40 and its ligand, CD154. *Cancer Res.* **57**, 2569–2574 (1997).
19. Rolink, A., Melchers, F. & Andersson, J. The SCID but not the RAG-2 gene product is required for the S mu-S epsilon heavy chain class switching. *Immunity* **5**, 319–330 (1996).
20. Grusby, M. J., Johnson, R. S., Papaioannou, V. E. & Glimcher, L. H. Depletion of CD4+ T cells in major histocompatibility complex class II-deficient mice. *Science* **253**, 1417–1420 (1991).
21. Shinde, S. *et al.* CD40L is important for induction of, but not response to, costimulatory activity. ICAM-1 as the second costimulatory molecule rapidly upregulated by CD40L. *J. Immunol.* **157**, 2764–2768 (1996).
22. Kitamura, D., Roes, J., Kuhn, R. & Rajewsky, K. A B cell-deficient mouse by targeted disruption of the membrane exon of the immunoglobulin mu chain gene. *Nature* **350**, 423–426 (1991).

23. Noelle, R. J. *et al.* A 39-kDa protein on activated T helper cells binds CD40 and transduces the signal for cognate activation of B cells. *Proc. Natl Acad. Sci. USA* **89**, 6550–6554 (1992).
24. Yang, Y. & Wilson, J. M. CD40 ligand-dependent T cell activation: requirement of B7-CD28 signaling through CD40. *Science* **273**, 1862–1864 (1996).
25. Steinman, R. M. The dendritic cell system and its role in immunogenicity. *Annu. Rev. Immunol.* **9**, 271–296 (1991).
26. Mayordomo, J. I. *et al.* Bone marrow-derived dendritic cells pulsed with synthetic tumour peptides elicit protective and therapeutic antitumour immunity. *Nature Med.* **1**, 1297–1302 (1995).
27. Kurts, C., Kosaka, H., Carbone, F. R., Miller, J. F. A. P. & Heath, W. R. Class I-restricted cross-presentation of exogenous self-antigens leads to deletion of autoreactive CD8+ T cells. *J. Exp. Med.* **186**, 239–245 (1997).
28. Kast, W. M. *et al.* Eradication of adenovirus E1-induced tumors by E1a-specific cytotoxic T lymphocytes. *Cell* **59**, 603–614 (1989).
29. Dialynas, D. P. *et al.* Characterization of the murine T cell surface molecule, designated L3T4, identified by the monoclonal antibody GK1.5: similarity of L3T4 to the human Leu-3/T4 molecule. *J. Immunol.* **131**, 2445–2451 (1983).

Acknowledgements. We thank P. Matzinger for discussions and for providing the B6I-A^b-knockout and μ MT mice used in this study; A. Rolink for the FGK45 hybridoma; J. Laman for the purified MR1 antibody; and F. Ossendorp, F. Koning and R. De Vries for critically reading the manuscript and for making suggestions for its improvement. This work was funded by grants from the Dutch Cancer Society and the European Community. R.E.M.T. is a fellow of the Royal Dutch Academy of Arts and Sciences.

Correspondence and requests for materials should be addressed to S.P.S. (e-mail: sps@liai.org).

Role of Mxi1 in ageing organ systems and the regulation of normal and neoplastic growth

Nicole Schreiber-Agus^{*†}, Yong Meng[†], Tin Hoang[†], Harry Hou Jr[†], Ken Chen[†], Roger Greenberg[†], Carlos Cordon-Cardo[‡], Han-Woong Lee^{†§} & Ronald A. DePinho[†]

Departments of ^{*}Molecular Genetics and [†]Microbiology and Immunology, Albert Einstein College of Medicine, Bronx, New York 10461, USA
[‡]Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA

Mxi1 belongs to the Mad (Mxi1) family of proteins, which function as potent antagonists of Myc oncoproteins^{1–4}. This antagonism relates partly to their ability to compete with Myc for the protein Max and for consensus DNA binding sites and to recruit transcriptional co-repressors^{4–6}. Mad(Mxi1) proteins have been suggested to be essential in cellular growth control and/or in the induction and maintenance of the differentiated state^{6,7}. Consistent with these roles, *mxi1* may be the tumour-suppressor gene that resides at region 24–26 of the long arm of chromosome 10. This region is a cancer hotspot, and mutations here may be involved in several cancers, including prostate adenocarcinoma^{8–10}. Here we show that mice lacking Mxi1 exhibit progressive, multisystem abnormalities. These mice also show increased susceptibility to tumorigenesis either following carcinogen treatment or when also deficient in Ink4a. This cancer-prone phenotype may correlate with the enhanced ability of several *mxi1*-deficient cell types, including prostatic epithelium, to proliferate. Our results show that Mxi1 is involved in the homeostasis of differentiated organ systems, acts as a tumour suppressor *in vivo*, and engages the Myc network in a functionally relevant manner.

To disrupt the *mxi1* open reading frame (ORF), we designed a positive/negative replacement-type vector that eliminates an exon required for the production of the two mouse Mxi1 isoforms⁴ (Fig. 1a). On transmission of the *mxi1*-null allele through the germ line, heterozygous intercrosses yielded all three genotypes (Fig. 1b) at a ratio close to the expected Mendelian distribution (relative ratios *mxi1*^{+/+} 1, *mxi1*^{+/-} 2.5, *mxi1*^{-/-} 0.9; *n* = 118). We used reverse transcription with polymerase chain reaction (RT-PCR), with a 5'-oligomer targeted to the deleted region, to confirm that the

[§]Present address: Department of Biochemistry and Molecular Biology, Ilchun Institute for Molecular Medicine, Seoul National University College of Medicine, Seoul, Korea.

mutant mice indeed had a disrupted *mxil* locus. RT-PCR produced an *mxil*-specific band in total tissue RNA from adult *mxil*^{+/+}, but not *mxil*^{-/-} mice (Fig. 1c). *mxil*^{-/-} mice were viable, fertile and outwardly healthy.

We performed an extensive histological survey of young and aging organs from wild-type and homozygous-null *mxil* mice. We focused particularly on organs that normally express high or sustained levels of *mxil* (for example, brain, spleen, kidney and liver) (N.S.-A. and R.A.D., unpublished observations) and on tissue types that are susceptible to tumorigenesis when a putative tumour suppressor is lost from the chromosome 10q24-26 region, to which *mxil* maps⁸⁻¹⁰ (for example, the spleen and thymus are susceptible to T-cell leukaemia, the prostatic epithelium to prostate cancer, and the brain to glioblastoma multiforme when the 10q24-26 region is mutated).

On gross inspection of the peritoneal cavity, the spleens of *mxil*^{-/-} mice were seen to be slightly enlarged in comparison to age-matched controls (data not shown). This macroanatomical finding correlated with a cellular expansion of the white pulp of *mxil*^{-/-} spleens, characterized initially by fusion of white-pulp follicles and followed by moderate to severe extramedullary haematopoietic changes (compare Fig. 2a, b). In some instances, lymphocytic precursor cells also showed dysplastic morphology. This hyperplasia/dysplasia did not translate into changes in the splenic subpopulation profiles, as determined by fluorescence-activated cell sorting (FACS) after staining for cell-surface differentiation markers (data not shown). Instead these changes may be preneoplastic in nature, as suggested by the finding that several older *mxil*^{-/-} mice, and none of the *mxil*^{+/+} mice, succumbed to malignant lymphomas (of B-cell origin) as they approached 1 year of age (Fig. 2c).

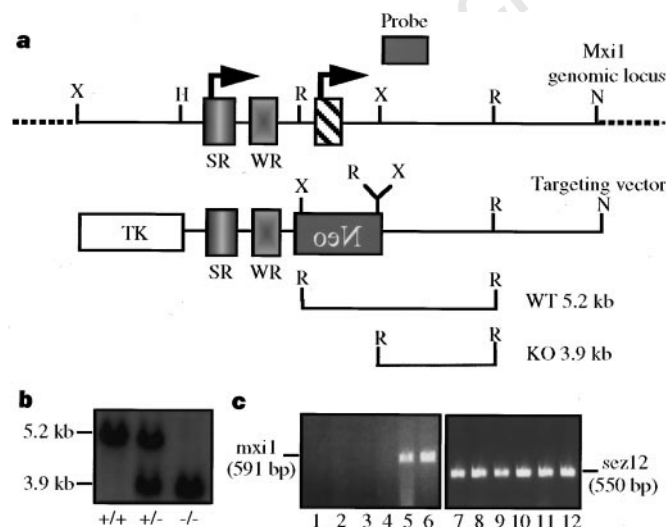


Figure 1 Generation of *Mxi1*-deficient mice. **a**, Genomic organization of the 5' region of the mouse *mxil* gene, showing the exons specific to *Mxi1*-SR (SR) and *Mxi1*-WR (WR) and their shared exon (striped), which is replaced by the neomycin-resistance (Neo) cassette after homologous recombination with the targeting vector. The locations of the probe used in **b** and the generated restriction fragments corresponding to wild-type (WT) and disrupted (KO) alleles are indicated. R, *EcoRI*; X, *XbaI*; N, *NotI*; H, *HindIII*; TK, thymidine kinase. **b**, Southern blot of *EcoRI*-digested tail DNA from the offspring of *mxil*^{+/+} intercrosses. **c**, Left, ethidium-bromide-stained products from RT-PCR experiments using *mxil* primers from the deleted shared exon and the 3' ORF. Right, control PCRs using oligomers to mouse *sez12* on the same first-strand syntheses. Lanes 1, 3, 7 and 9, *mxil*^{-/-} adult brain; lanes 2, 4, 8 and 10, *mxil*^{-/-} adult kidney; lanes 5 and 11, wild-type adult brain; lanes 6 and 12, wild-type adult kidney.

These splenic abnormalities are consistent with a role for members of the Mad(*Mxi1*) family in haematopoietic growth control; such a role is indicated by their upregulated expression during induced differentiation in cell culture (reviewed in refs 6, 7). These alterations are also reminiscent of those that occur in the E μ -myc transgenic model¹¹, in which susceptibility to B-cell versus T-cell lymphomas is affected by genetic background^{12,13}. As the mixed genetic background of the *Mxi1*-deficient mouse may also predispose to lymphoma type, we cannot rule out the possibility that, in some cases, the absence of *Mxi1* may contribute to T-cell leukaemias resulting from 10q24-26 loss¹⁴.

Analysis of the genitourinary system of *mxil*^{-/-} mice showed that there were profound and progressive degenerative changes in the cortical region of the kidney. In the younger mice, the first renal alterations involved cytoplasmic vacuolization of proximal tubular epithelial cells (compare Fig. 2d, e). These proximal tubular changes resemble the focal dilations observed in the precystic phase of polycystic kidney disease (PKD), a condition that has been linked both clinically and experimentally to c-Myc dysregulation^{15,16}. In *mxil*^{-/-} mice aged 7 months or more, we saw atrophy of proximal tubules and foci of regenerating proximal convoluted tubules (data not shown). Focal atrophic changes in the glomeruli also occurred, with partial to total disappearance of the glomerular tufts and dilation of the Bowman's space taking place (Fig. 2f). The end result of these alterations appears to be the production of cystic spaces similar to those described in human glomerulocystic diseases of the kidney, diseases that have no known pathogenetic basis (reviewed in ref. 17). Our results indicate that, like overexpression of *Myc*, loss of *Mxi1* function contributes to renal cystic disease, further implicating the *Myc* pathway in the normal kidney developmental programme.

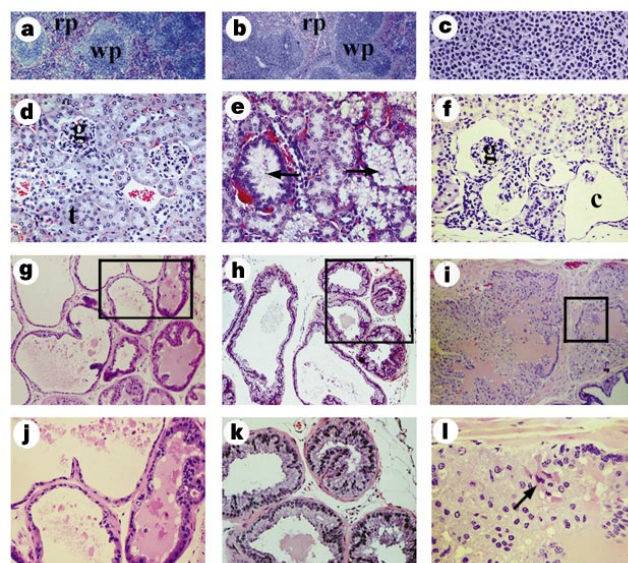


Figure 2 Histological survey of highly affected tissues in *mxil*^{-/-} mice. **a-c**, Spleen. Haematoxylin and eosin-stained spleen from 11-month-old wild-type (**a**) and *mxil*^{-/-} (**b**) mice. **c**, A malignant lymphoma arising spontaneously in a 12-month-old *mxil*^{-/-} mouse. Rp, red pulp; wp, white pulp. This histopathology was seen in 8 out of 24 *mxil*^{-/-} and 0 out of 11 *mxil*^{+/+} spleens examined. **d-f**, Kidney. **d**, Wild-type kidney, 1 yr of age. **e, f**, *mxil*^{-/-} kidney, 1 yr of age. Arrows in **e** point to vacuolated tubules. G, glomerulus; t, tubule; c, cyst. This histopathology was seen in 15 out of 24 *mxil*^{-/-} and 1 out of 11 *mxil*^{+/+} kidneys examined. **g-i**, Prostate. **g, j**, Wild-type prostate, 1 yr of age. **h, k**, *mxil*^{-/-} prostate, 7 months of age. **i, l**, *mxil*^{-/-} prostate, 1 yr of age. Areas of detail shown in **j-l** are boxed in **g-i**. Arrow in **l** points to a representative mitotic figure. This histopathology was observed in 6 out of 11 *mxil*^{-/-} and 1 out of 4 *mxil*^{+/+} prostates examined. Original magnification: **a, b**, $\times 40$; **g-i**, $\times 100$; **d-f, j, k**, $\times 200$; **c, l**, $\times 400$.

We also observed histopathological changes in the prostate glands of the *mxil*-null mice. These changes became more noticeable with advancing age (compare Fig. 2g, h, j, k). Specifically, older (~1-yr-old) *mxil*^{-/-} mice consistently showed an increase in the total number of prostatic glands (atrophic + acinar) compared with age-matched wild-type mice (average 98.4 ± 18.7 glands total for *mxil*^{-/-} prostates (*n* = 9) versus 62 ± 12.1 for *mxil*^{+/+} prostates (*n* = 4)). Moreover, *mxil*-deficient prostates often contained microscopic foci of enlarged and complex glandular structures, some of them occupied by hypercellular acini and dysplastic cells, with occasional mitotic figures in epithelial cells located in supra-basal layers of the acini (Fig. 2i, l). In the 11 *mxil*^{-/-} prostatic tissues analysed, there was no overt neoplastic transformation, indicating that Mx1 deficiency alone may lead solely to a preneoplastic condition (see below).

The hyperplastic changes seen in the splenic white pulp and in the prostate indicate a role for Mx1 in cellular growth control. We investigated this role further by examining the consequences of Mx1 deficiency on the growth behaviour of specific mouse cell types, including lymphocytes, prostatic epithelium, and embryonic fibroblasts. First, to evaluate the proliferative capacity of Mx1-deficient and control lymphocytes, we cultured age-matched splenocytes derived from mice ranging from 1–4 months in age in the presence or absence of T-cell- or B-cell-specific mitogens (anti-CD3 plus anti-CD28 antibodies or bacterial lipopolysaccharide (LPS), respectively). Although the extent of ³H-thymidine incorporation (a measure of cell proliferation) was comparable in the untreated samples, it was consistently increased in the *mxil*^{-/-} cultures after stimulation with anti-CD3 plus anti-CD28 antibodies (Fig. 3A) but not with LPS (data not shown). To determine the basis for the increased ³H-thymidine incorporation, we assayed the cell-cycle profile and degree of apoptosis following anti-CD3/CD28 stimulation of wild-type and *mxil*^{-/-} splenocytes. In the *mxil*^{-/-} samples there was an increase in exit from the G1 phase of the cell cycle that was proportional to the increase in ³H-thymidine incorporation (Fig. 3A, right panels; overall 10% decrease in G1 population of anti-CD3/CD28-treated *mxil*^{-/-} cells). There were no changes in the percentage of apoptotic (annexin-positive) cells (data not shown). These results indicate that Mx1 may function normally to control T-cell growth in response to mitogenic stimuli by regulating exit from G1 phase. A similar *in vivo* role for the related Mad1 protein

has been proposed based on the finding that cluster-forming colonies from Mad1-deficient mice are less able to withdraw from the cell cycle before entry into terminal differentiation¹⁸.

We next assessed proliferation in the prostatic epithelium by immunohistochemical analyses of formalin-fixed, paraffin-embedded sections through the prostate of Mx1-deficient and age-matched wild-type males, aged 7 months and more, using an anti-Ki67 antibody. The Ki67 protein is a reliable marker of cellular proliferation in rodent cells¹⁹ and easily identified cycling cells in positive-control lymph-node sections from wild-type mice (Fig. 3Bb). On the prostatic sections, there was positive anti-Ki67 immunostaining in an average of 7.3 out of 1,103 *mxil*^{-/-} (*n* = 9 mice) and 0.25 out of 1,096 *mxil*^{+/+} (*n* = 4 mice) prostatic epithelial cells examined in several high-power microscopic fields per sample (Fig. 3Ba, c, d). These results indicate that the prostatic hyperplasia observed histologically (see above) probably results from aberrant S-phase entry due to Mx1 deficiency.

Finally, we assayed the proliferative properties of cultured Mx1-deficient and wild-type mouse embryonic fibroblasts (MEFs) by growth curves and by flow cytometry of exponentially growing cultures. There were modest yet consistent increases in cell numbers and doubling time and in the S-phase population of *mxil*^{-/-} MEFs (data not shown). As a further measure of proliferation potential, we compared the ability of *mxil*^{-/-} and *mxil*^{+/+} MEFs to produce visible colonies following low-density seeding. Mx1-deficient cells exhibited enhanced ability to generate actively growing colonies of clonal origin (Fig. 3C). This growth advantage of Mx1-deficient fibroblasts may provide the basis for our observations in transformation assays, in which foci formation by *myc/Ras* was increased by three- to fivefold in *mxil*^{-/-} MEFs as compared with in wild-type controls (data not shown).

As loss of Mx1 function allowed transformation in cell culture and leads to spontaneous lymphomas at low frequency and long latency (see above), we further assessed the tumour suppressor activity of Mx1 *in vivo* using two approaches that accelerate tumorigenesis. In the first approach, the tumour susceptibility of wild-type and *mxil*^{-/-} mice arising from heterozygous intercrosses was compared following weekly topical application of the carcinogen 9,10-dimethyl-1,2-benzanthracene (DMBA) (reviewed in ref. 20). Over an observation period of 25 weeks, 75% of the *mxil*^{+/+} mice (*n* = 28) but only 35% of the *mxil*^{-/-} mice (*n* = 23) survived

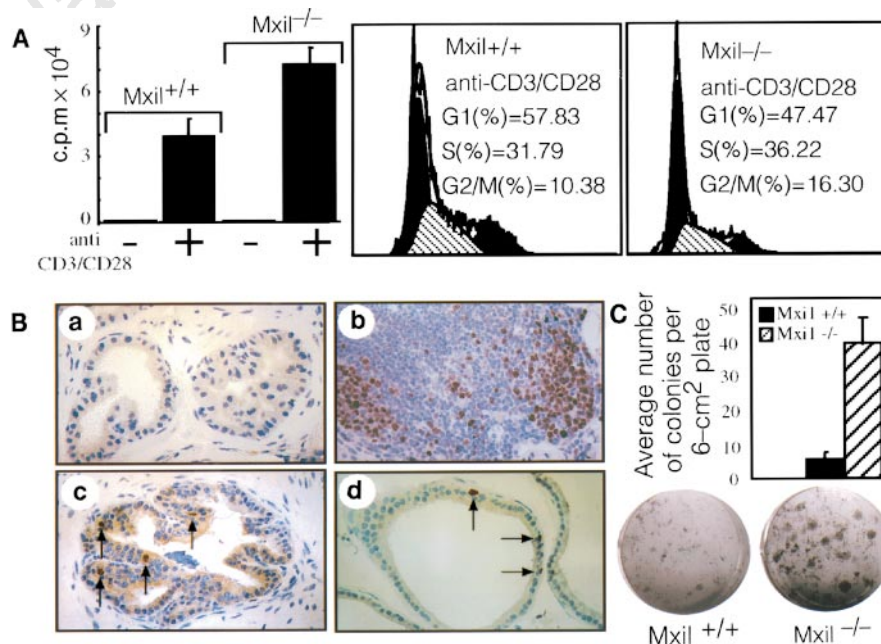


Figure 3 Mx1-dependent growth control in diverse cell types. **A**, Splenocytes. Left panel, comparison of splenocyte proliferation as measured by ³H-thymidine incorporation following 48 h incubation with or without exposure to the T-cell-specific mitogens anti-CD3/anti-CD28 antibodies. Error bars represent standard deviations for samples tested in quadruplicate. Right panels, cell-cycle profiles of mitogen-treated cultures from a separate experiment after labelling with propidium iodide and analysis by FACS. **B**, Prostatic epithelium. Immunohistochemical analysis of proliferation using an anti-Ki67 antibody on wild-type (a) and *mxil*-deficient (c, d) prostatic epithelium, and on wild-type lymph nodes as a positive control (b). Arrows point to immunoreactivity in the *mxil*^{-/-} nuclei. Original magnification, ×400. **C**, Embryonic fibroblasts. The histogram represents the average number of colonies arising 8 days after low-density seeding of wild-type or *mxil*^{-/-} MEFs. Representative Giemsa-stained plates from a separate low-density experiment are shown below.

this treatment and remained free of clinically apparent tumours (Fig. 4a; $P = 0.004$). The major tumour types seen were squamous cell carcinoma of the skin and malignant lymphoma, cancers of two cell types that are known to express *mxil* (N.S.-A. and R.A.D., unpublished observations).

The second approach evaluated whether loss of Mxi1 function affects the incidence or latency of tumours arising in the Ink4a-deficient mouse, which develops spontaneous tumours at an early age, the predominant tumour types being fibrosarcoma and lymphoma²¹. Over an observation period of 30 weeks, 80% of the *ink4a*^{-/-} mice ($n = 15$) and only 44% of mice double null for both *ink4a* and *mxil* ($n = 16$) survived and remained tumour-free (Fig. 4b; $P = 0.038$). The spectrum of tumour types seen in the *mxil*^{-/-}, *ink4a*^{-/-} population resembled that seen with Ink4a deficiency alone (data not shown). A functional relationship between *myc* deregulation and the Ink4a tumour-suppressor pathway was indicated previously by the ability of overexpressed p16^{INK4a} or p19^{ARF} to suppress Myc/Ras cooperativity in the rat embryonic fibroblast transformation assay^{22,23}. Here, the ability of loss of function of Mxi1 to enhance the cancer-prone condition of the Ink4a-deficient mouse provides *in vivo* confirmation of this relationship.

Our phenotypic analysis of the *mxil*^{-/-} mouse indicates that a member of the Mad(Mxi1) family is vital in the regulation of normal cellular growth, the maintenance of mature organ systems, and the suppression of cancer. The striking concordance in phenotypes resulting from Myc overexpression (reported previously) and from Mxi1 deficiency (for example, cancer and PKD) lends support at the organismal and cellular levels for the widely held model that Myc and Mxi1 antagonize each other in regulating many of the same biological processes, by engaging common gene targets (reviewed in refs 6, 7). One important corollary of the late-onset and progressive *mxil*-null phenotype is that the long-term stability of some terminally differentiated tissues (such as the genitourinary and haematopoietic systems) requires not only downregulation of Myc expression but also sustained and active Mxi1-mediated repression. The more pronounced defects associated with Mxi1 deficiency in the mouse, in comparison with those recently described for the *mad1*^{-/-} mouse¹⁸, may relate to tissue specificity of Mxi1 versus Mad expression, to lack of functional redundancy in the affected tissues, or to differences in genetic background, among other possibilities.

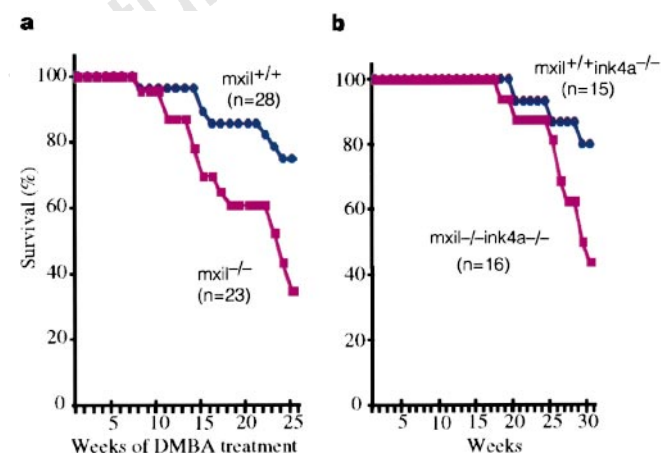


Figure 4 Mxi1-mediated tumour suppression *in vivo*. **a**, Survival curves of DMBA-treated wild-type and *mxil*^{-/-} mice derived from heterozygous intercrosses. **b**, Survival curves of mice doubly null for both *mxil* and *ink4a* or null for *ink4a* alone; these mice were derived from crosses between mating pairs heterozygous-null for *mxil* and homozygous-null for *ink4a*. For all informative cases, the cause of death was the development of a clinically apparent and histologically confirmed tumour. During the observation period of 25–30 weeks, all untreated wild-type mice (>100) remained healthy and tumour-free.

Our discovery of ectopic proliferation in the Mxi1-deficient prostatic epithelium defines a role for Mxi1 in normal prostate biology. It is unknown whether Mxi1 is the 10q24-26-located tumour suppressor that is implicated in prostate carcinogenesis; a newly described 10q24-26-located candidate tumour suppressor gene, *PTEN*, has been found to be mutated in 4 out of 4 prostate cancer cell lines tested²⁴. Is *MXI1* mutated in primary tumours? Early studies yielded conflicting results^{25–27}, but in a recent study submicroscopic loss of a single *MXI1* allele was seen in 53% of the primary tumours examined²⁸, with many of these samples exhibiting mutational inactivation of the retained *MXI1* allele. In addition, a causative link between the Myc superfamily and prostatic carcinogenesis has been proposed, on the basis of the observations that Myc gene amplification was shown to be a potential marker of prostate carcinoma progression²⁹ and that retroviral introduction of *myc* plus *ras* into the urogenital sinus resulted in poorly differentiated prostate cancers with high penetrance³⁰. The idea that additional genetic and epigenetic events may be needed to cooperate with Mxi1 deficiency in generating prostate adenocarcinoma concurs with the fact that human prostate cancer is a latent disease of the elderly. The Mxi1-knockout mouse provides an *in vivo* model system for the identification of these cooperating genetic lesions. □

Methods

Gene targeting. The mouse *mxil* 5' genomic region was isolated by screening a 129/Sv mouse genomic library with the 5' portion of the *mxil* complementary DNA⁴. In the targeting construct, pMxi1KO, we used the mutant neomycin-resistance and the herpes simplex virus thymidine kinase genes under the controls of PGK promoters. With incorporation of the null allele, the initiator methionine of the Mxi1-WR isoform is eliminated. Additionally, the 133-base-pair size of the deleted exon (not divisible by 3) ensures that putative splicing events involving the SR exon would not maintain a continuous ORF from the Sin3-interacting domain through to the bHLH/LZ domain (Fig. 1a). pMxi1KO was electroporated into WW6 embryonic stem cells²¹, and we selected clones using G418 (150 µg ml⁻¹ active component) and 2 µM ganciclovir. *Mxi1*^{+/+} embryonic stem cell clones, identified by Southern blot analysis using a probe outside the 5' homology arm, were used to generate germline chimaeras. RT-PCR was performed with the *mxil*-specific primers 5'-GCAACACCAGC ACTGCCAAC-3' and 5'-AGGAGACTGCATCATGAACC-3'. The mouse *sez12* control primers were a gift from B. Saint-Jore and A. Skoultschi.

Histology and immunohistochemistry. Following microwave antigen retrieval, sections were blocked and incubated with rabbit primary anti-Ki67 purified antiserum (Vector, Burlingame, CA) at 1:10,000 dilution (or with a rabbit preimmune serum as a negative control, at the same working dilution) overnight at 4°C. Secondary reactions with goat anti-rabbit biotinylated antibodies (1:1,000 dilution, Vector) were for 30 min at room temperature; these were followed with avidin-biotin-peroxidase complexes (1:25 dilution, Vector) for 30 min at room temperature. Diaminobenzidine was used as the final chromogen and haematoxylin as the nuclear counterstain. Nuclear expression of Ki67 (ref. 19) was assessed by counting at least 1,000 cells from high-power-microscopy fields (×400 magnification).

Analysis of Mxi1-deficient embryonic fibroblasts and splenocytes. Colony formation and focus assays of embryonic fibroblasts were performed as described²¹. For the ³H-thymidine-incorporation assays, we resuspended 125 µl dispersed splenocytes at 2 × 10⁶ per ml in RPMI medium and plated them into round-bottomed 96-well plates. The mitogens used were LPS (4 µg per well, Sigma), purified hamster anti-mouse CD3 (10 ng ml⁻¹, Pharmingen), and purified hamster anti-mouse CD28 (10 ng ml⁻¹, Pharmingen). At 24, 48 and 72 h after addition of mitogen, cultures were pulse-labelled for 4–6 h with ³H-thymidine (diluted 1:10) and collected onto glass fibre filter strips. We then determined the incorporated c.p.m. for each cell lysate. Samples were processed in quadruplicate and their averages and standard deviations were calculated by dedicated software. For analysis of cell cycle and apoptotic profiles, 5 ml splenocytes resuspended at 2 × 10⁶ per ml as above were plated into 6-well plates in the presence or absence of anti-CD3/CD28 antibodies. After 48 h, cells

were stained with either propidium iodide alone (after citric acid treatment) or with annexin-V-fluos (Boehringer) in a HEPES buffer containing propidium iodide, and were analysed by FACScan (Becton Dickinson, San Jose). LYSIS II was used for instrument control, data acquisition, and analysis.

DMBA tumorigenic treatments. DMBA treatments consisted of topical applications of 50 μ l of 1.25 mg DMBA per ml acetone to the dorsal surface, beginning at postnatal day 2–5 and continuing weekly thereafter. Mice were monitored regularly and killed when overall health was compromised. A full histopathological survey was conducted on the first 8–10 mice in each cohort; thereafter, detailed pathological analysis was restricted to the site of tumour involvement and organs of potential metastatic involvement (for example, lymphoid organs, liver, kidney and lung). For the two *in vivo* tumour studies, Kaplan–Meier estimates of the survival probability were calculated and comparisons between the groups were made using the log rank test.

Received 26 January; accepted 18 March 1998.

1. Ayer, D. E., Kretzner, L. & Eisenman, R. N. Mad: a heterodimeric partner for Max that antagonizes Myc transcriptional activity. *Cell* **72**, 211–222 (1993).
2. Hurlin, P. J. et al. Mad3 and Mad4: novel Max-interacting transcriptional repressors that suppress c-myc dependent transformation and are expressed during neural and epidermal differentiation. *EMBO J.* **14**, 5646–5659 (1995).
3. Zervos, A. S., Gyuris, J. & Brent, R. Mx1, a protein that specifically interacts with Max to bind Myc-Max recognition sites. *Cell* **72**, 223–232 (1993).
4. Schreiber-Agus, N. et al. An amino-terminal domain of Mx1 mediates anti-Myc oncogenic activity and interacts with a homolog of the yeast transcriptional repressor Sin3. *Cell* **80**, 777–786 (1995).
5. Ayer, D. E., Lawrence, Q. A. & Eisenman, R. N. Mad-Max transcriptional repression is mediated by ternary complex formation with mammalian homologs of yeast repressor Sin3. *Cell* **80**, 767–776 (1995).
6. Schreiber-Agus, N. & DePinho, R. A. Repression by the Mad(Mx1)–Sin3 complex. *Bioessays* (in the press).
7. Henriksson, M. & Luscher, B. Proteins of the Myc network: essential regulators of cell growth and differentiation. *Adv. Cancer Res.* **68**, 109–182 (1996).
8. Edelhoff, S. et al. Mapping of two genes encoding members of a distinct subfamily of MAX interacting proteins: MAD to human chromosome 2 and mouse chromosome 6, and MX11 to human chromosome 10 and mouse chromosome 19. *Oncogene* **9**, 665–668 (1994).
9. Shapiro, D. N. et al. Assignment of the human MAD and MX11 genes to chromosomes 2p12–p13 and 10q24–q25. *Genomics* **23**, 282–285 (1994).
10. Wechsler, D. S. et al. Localization of the human Mx11 transcription factor gene (MX11) to chromosome 10q24–q25. *Genomics* **21**, 669–672 (1994).
11. Adams, J. M. et al. The c-myc oncogene driven by immunoglobulin enhancers induces lymphoid malignancy in transgenic mice. *Nature* **318**, 533–538 (1985).
12. Harris, A. W. et al. Transgenic mouse models for hematopoietic tumorigenesis. *Curr. Top. Microbiol. Immunol.* **141**, 82–93 (1988).
13. Yukawa, K. et al. Strain dependency of B and T lymphoma development in immunoglobulin heavy chain enhancer (E μ)–myc transgenic mice. *J. Exp. Med.* **170**, 711–726 (1989).
14. Raimondi, S. C. et al. Cytogenetics of childhood T-cell leukemia. *Blood* **72**, 1560–1566 (1988).
15. Cowley, B. D. Jr, Chadwick, L. J., Grantham, J. J. & Calvet, J. P. Elevated proto-oncogene expression in polycystic kidneys of the C57BL/6J (cpk) mouse. *J. Am. Soc. Nephrol.* **1**, 1048–1053 (1991).
16. Trudel, M., D'Agati, V. & Constantini, F. C-myc as an inducer of polycystic kidney disease in transgenic mice. *Kidney Int.* **39**, 665–671 (1991).
17. Dedeglu, I. O. et al. Spectrum of glomerulocystic kidneys: a case report and review of the literature. *Ped. Path. Lab. Med.* **16**, 941–949 (1996).
18. Foley, K. P. et al. Targeted disruption of the MYC antagonist MAD1 inhibits cell cycle exit during granulocyte differentiation. *EMBO J.* **17**, 774–785 (1998).
19. Gerlach, C. et al. Ki-67 immunoreactivity is a robust marker of proliferative cells in the rat. *Lab. Invest.* **77**, 697–698 (1997).
20. Hecker, E. et al. International symposium: skin carcinogenesis in man and in experimental models. *J. Cancer Res. Clin. Oncol.* **118**, 321–328 (1992).
21. Serrano, M. et al. Role of the INK4a locus in tumor suppression and cell mortality. *Cell* **85**, 27–37 (1996).
22. Serrano, M. et al. Inhibition of ras-induced proliferation and cellular transformation by p16INK4. *Science* **267**, 249–252 (1995).
23. Pomerantz, J. et al. The INK4a tumor suppressor gene product, p19ARF, interacts with MDM2 and neutralizes MDM2's inhibition of p53. *Cell* **92**, 713–723 (1998).
24. Li, J. et al. PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. *Science* **275**, 1943–1947 (1997).
25. Eagle, L. R. et al. Mutation of the MX11 gene in prostate cancer. *Nature Genet.* **9**, 249–255 (1995).
26. Kawamata, N. et al. Point mutations of the Mx11 gene are rare in prostate cancers. *Prostate* **29**, 191–193 (1996).
27. Gray, J. C. et al. Loss of the chromosomal region 10q23–25 in prostate cancer. *Cancer Res.* **55**, 4800–4803 (1995).
28. Prochownik, E. V. et al. Commonly occurring loss and mutation of the MX11 gene in prostate cancer. *Genes Chromosom. Cancer* (in the press).
29. Jenkins, R. B., Qian, J., Lieber, M. M. & Bostwick, D. G. Detection of c-myc oncogene amplification and chromosomal anomalies in metastatic prostatic carcinoma by fluorescence in situ hybridization. *Cancer Res.* **57**, 524–531 (1997).
30. Thompson, T. C. et al. Transgenic models for the study of prostate cancer. *Cancer* **71**, 1165–1171 (1993).

Acknowledgements. We thank A. Skoultschi and members of the DePinho lab for comments; D. Gebhardt, for assistance and advice with FACS studies; L. Chin for graphic-art assistance; C. J. Chang for statistical analysis; and J. Lauridsen for administrative assistance. H.-W.L. and Y.M. were supported by NIH training grants. R.G. was supported by an MSTP training grant. N.S.-A. is a Special Fellow of the Leukemia Society of America. C.C.-C. is supported by grants from the NIH. R.A.D. is supported by grants from the NIH and by the Irma T. Hirsch Award. We acknowledge support from the Albert Einstein Cancer Center Core.

Correspondence and requests for material should be addressed to R.A.D. (e-mail: depinho@acom.yu.edu).

A cytokinesis checkpoint requiring the yeast homologue of an APC-binding protein

Li Muhua, Neil R. Adames, Michael D. Murphy, Colleen R. Shields & John A. Cooper

Department of Cell Biology and Physiology, Washington University School of Medicine, St Louis, Missouri 63110, USA

Checkpoint controls ensure that events of the cell-division cycle are completed with fidelity and in the correct order. In budding yeast with a mutation in the motor protein dynein, the mitotic spindle is often misaligned and therefore slow to enter the neck between mother cell and budding daughter cell. When this occurs, cytokinesis (division of the cytoplasm into two) is delayed until the spindle is properly positioned¹. Here we describe mutations that abolish this delay, indicating the existence of a new checkpoint mechanism. One mutation lies in the gene encoding the yeast homologue of EB1, a human protein that binds the adenomatous polyposis coli (APC) protein, a tumour suppressor. EB1 is located on microtubules of the mitotic spindle and is important in spindle assembly. EB1 may therefore, by associating with microtubules, contribute to the sensor mechanism that activates the checkpoint. Another mutation affects Stt4, a phosphatidylinositol-4-OH kinase. Cold temperature is an environmental stimulus that causes misalignment of the mitotic spindle in yeast and appears to activate this checkpoint mechanism.

Yeast lacking dynein show a delay in the cell cycle before cytokinesis; this delay allows the cell to correct defects in mitotic spindle orientation¹. The degree of the defect in mitotic spindle orientation and the delay in the cell cycle necessary for correction vary. About two-thirds of the cells show mild defects with short delays of ~5 min, and about one-third show more severe defects with longer delays of ~30 min. The existence of this delay suggested that a new cell-cycle checkpoint may exist. To investigate the existence of such a checkpoint, we did a genetic screen in *Saccharomyces cerevisiae*, searching for mutations that might abolish the cell-cycle delay.

We screened for mutations that are lethal in combination with the *act5* mutation. Act5 is necessary for dynein function². Two mutants, *stt4-7* and *yeb1-29*, caused a loss of the cell-cycle delay (see below). Using tetrad analysis we found that *stt4* and *yeb1* were lethal in combination with other mutations affecting dynein function, *dnc1* (ref. 1) and *jnm1* (ref. 3). We cloned *STT4* and *YEB1* by complementation. Linkage analysis showed that the mutations resided in the cloned genes. *Stt4* is a phosphatidylinositol-4-OH kinase⁴. *YEB1* encodes the yeast homologue of a human protein, EB1 (ref. 5). Human EB1 binds the carboxy-terminal region of the tumour suppressor APC (for adenomatous polyposis coli) protein⁵; this may be essential for APC's tumour-suppressing function⁶.

To determine whether the cell-cycle delay induced by loss of Act5 was caused by a checkpoint control mechanism, we examined how the different mutations affected this delay (Fig. 1a). A temperature-sensitive *act5* allele induced the defect in spindle orientation and nuclear migration. The level of penetrance of this phenotype observed here is similar to that observed in the dynein-null mutant *dnc1* (ref. 1). We also constructed a complete disruption of the *YEB1* gene. A *yeb1Δ* haploid was viable and similar in phenotype to the original mutant, *yeb1-29*, as described below.

First, we compared *YEB1* wild-type and *yeb1Δ* mutant cells that had temperature-sensitive mutations in *act5* in time-lapse movies. These movies revealed the cell-cycle delay induced by misalignment

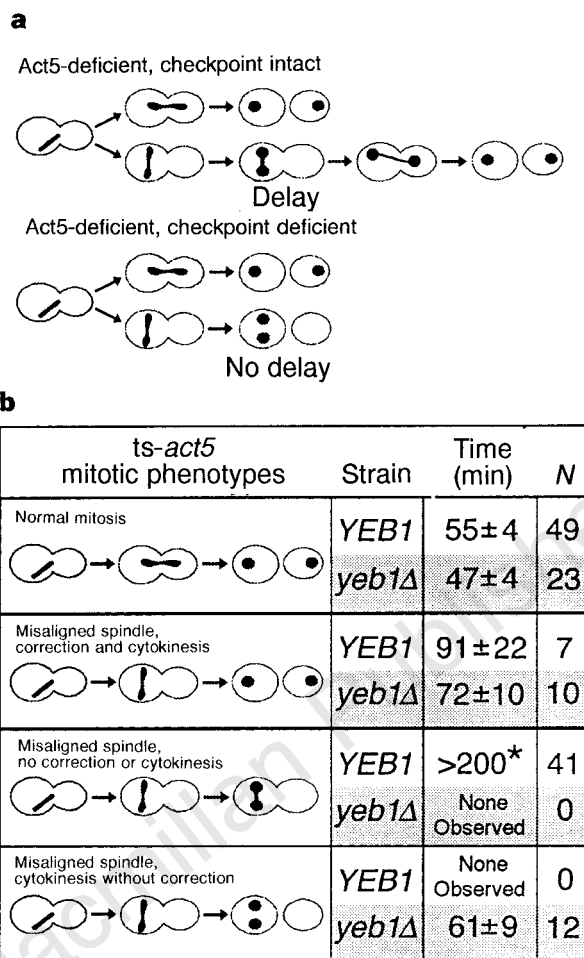


Figure 1 Movie assay for cell-cycle delay. **a**, Possible outcomes of mitosis in *act5* mutant cells. The upper of each pair of diagrams shows a normal mitosis. The lower shows a mitosis with a misaligned spindle. **b**, Experimental results. Time from spindle elongation to cell separation was measured in (temperature-sensitive) *ts-act5* *YEB1* and *ts-act5* *yeb1Δ* cells. Groups drawn at left include cells with properly aligned spindles, cells with misaligned spindles that corrected

these spindles, and cells with misaligned spindles that did not correct. *N*, number of cells. Some *ts-act5* *YEB1* cells did not correct their misorientated spindles (asterisk) and cell separation was not observed. These movies ended at 200–320 min. A time to cell separation of 200 min is significantly different from that of 55 min for normally dividing *ts-act5* *YEB1* cells, as is the value of 91 min for corrected misaligned *ts-act5* *YEB1* cells ($P < 0.01$).

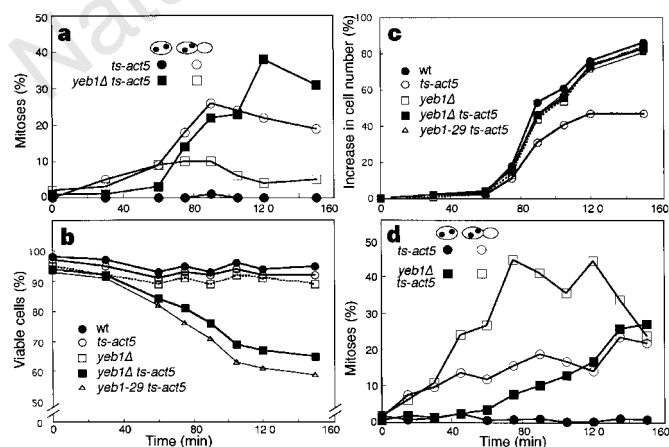


Figure 2 Synchronized cell assay for cell-cycle delay. Synchronization was achieved with hydroxyurea and α -factor in **a–c**, and with α -factor alone in **d**. **a**, **d**, Percentage of mitoses that produce binucleate large-budded cells and binucleate unbudded cells against time following release from arrest. **a**, One representative experiment. Two independent experiments gave similar results. **b**, **c**, Viability (**b**) and cell number (**c**) for these and additional control strains are plotted against time. Table 1 shows extra data from experiments in which hydroxyurea was used. Wt, wild type; ts, temperature sensitive.

of the mitotic spindle directly (Fig. 1b). When the mitotic spindle was orientated correctly, the time from spindle elongation to cell separation was ~50 min, with no delay. When the mitotic spindle was misaligned, some cells in both wild-type and mutant strains corrected their misaligned spindle and completed cytokinesis normally. *YEB1* wild-type cells exhibited delays before cytokinesis. Moreover, *YEB1* wild-type cells that did not realign their spindles halted in the cell cycle and did not proceed to cytokinesis.

In contrast, *yeb1Δ* mutant cells proceeded promptly with cell division even when the bud did not contain a nucleus, generating a binucleate mother and an anucleate daughter. Therefore, *YEB1* wild-type cells delay cell division to correct defects in spindle orientation, whereas *yeb1Δ* mutants proceed to cytokinesis even if the daughter does not contain a nucleus.

Second, we synchronized cells at the same point in the cell cycle by arresting the cell cycle using hydroxyurea. These cells were shifted to the temperature that is restrictive for growth of the *act5* temperature-sensitive mutant and released from the cell-cycle arrest. Delay before cell division caused accumulation of cells with large buds with both nuclei in the mother, and loss of delay caused appearance of binucleate unbudded cells (Fig. 1a). In temperature-sensitive-*ts-act5* *YEB1* cells, large-budded cells with both nuclei in the mother accumulated at 20–25% of mitoses (Fig. 2a and Table 1). The cell-cycle delay was also reflected in the increase in cell number, which

Table 1 Cell-cycle-delay assay

Time after hydroxyurea release (min)	Strain	Unbudded cells			Budded cells		
		No nucleus	One nucleus per cell	Two nuclei	Both nuclei in mother	One nucleus in mother, one in bud	One nucleus in mother, none in bud
0	WT	0	11	0	1	1	87
	<i>act5-ts</i>	0	10	0	0	0	88
	<i>yeb1Δ</i>	0	9	0	0	0	91
	<i>yeb1Δact5-ts</i>	0.5	8	1	2	1	87
	<i>yeb1-29act5-ts</i>	0.5	8	1.5	2.5	1.5	81
30	WT	0	14	0	0	7	79
	<i>act5-ts</i>	0	13	0	4	5	78
	<i>yeb1Δ</i>	0	11	0.5	0.5	6	81
	<i>yeb1Δact5-ts</i>	0.5	11	1	3	5	79
	<i>yeb1-29act5-ts</i>	1	9	1	2	6	81
60	WT	0	15	0	1	26	58
	<i>act5-ts</i>	0	15	0	8	23	54
	<i>yeb1Δ</i>	0	18	1	1	20	60
	<i>yeb1Δact5-ts</i>	1	17	3	8	19	52
	<i>yeb1-29act5-ts</i>	3	14	4	9	21	49
75	WT	0	36	0	1	45	18
	<i>act5-ts</i>	0	32	0	14	35	19
	<i>yeb1Δ</i>	0	34	1	1	41	23
	<i>yeb1Δact5-ts</i>	8	21	11	8	34	18
	<i>yeb1-29act5-ts</i>	9	22	12	7	35	15
90	WT	0	66	1	1	24	8
	<i>act5-ts</i>	0	44	1	19	25	11
	<i>yeb1Δ</i>	0	59	1	1	26	13
	<i>yeb1Δact5-ts</i>	7	39	15	7	23	9
	<i>yeb1-29act5-ts</i>	10	35	19	6	20	10
105	WT	0	77	0	1	16	6
	<i>act5-ts</i>	0	57	0	16	23	4
	<i>yeb1Δ</i>	1	69	1	1	21	7
	<i>yeb1Δact5-ts</i>	8	55	16	4	16	11
	<i>yeb1-29act5-ts</i>	9	48	18	4	13	8
120	WT	0	84	1	0	13	2
	<i>act5-ts</i>	0	63	0	14	19	3
	<i>yeb1Δ</i>	0	83	1	1	9	6
	<i>yeb1Δact5-ts</i>	15	51	21	2	8	3
	<i>yeb1-29act5-ts</i>	13	55	19	3	6	4
150	WT	0	89	1	0	6	4
	<i>act5-ts</i>	0	77	0	11	9	3
	<i>yeb1Δ</i>	0	87	1	1	8	3
	<i>yeb1Δact5-ts</i>	11	60	17	3	5	4
	<i>yeb1-29act5-ts</i>	12	63	16	3	3	3

Data in the last six columns indicate the percentages of different cell forms resulting over time in the indicated strains. A subset of these data is plotted in Fig. 2. WT, wild type; ts, temperature sensitive.

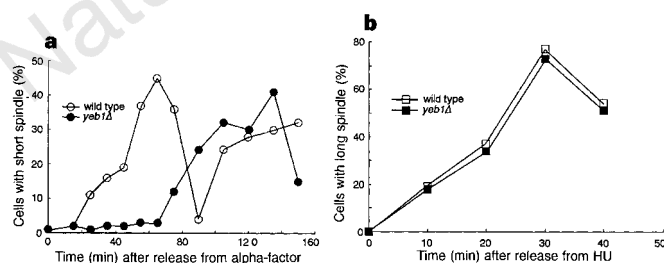


Figure 3 Spindle formation and elongation. **a**, Formation of the short pre-anaphase spindle over time. Cells were synchronized with α -factor and released. **b**, Elongation of the spindle, from short to long, over time. Cells were synchronized with hydroxyurea and released. More than 200 cells were counted for each time point.

was low in temperature-sensitive-*act5* cells relative to *ACT5* cells (Fig. 2c) because of a lack of cytokinesis.

In contrast, temperature-sensitive-*act5 yeb1Δ* cells showed no delay before cell division, as is shown by the increased number (~40%) of binucleate unbudded cells (Fig. 2a and Table 1). The *yeb1-29* (Table 1) and *stt4-7* mutants behaved similarly. Loss of cell-

cycle delay in *yeb1* mutants was also reflected in the increase in cell number, which returned to normal (Fig. 2c). Synchronization with α -factor instead of hydroxyurea gave similar results, albeit with less synchrony (Fig. 2d).

When a cell-cycle delay is abolished by a loss-of-function mutation, a checkpoint control mechanism is the likely cause⁷ of the delay. Complete disruption of *YEB1* caused a loss of the delay before cytokinesis in *act5* mutants; therefore, this delay is probably due to a checkpoint mechanism.

In the synchronized cell experiment, temperature-sensitive-*act5 yeb1* mutants experienced a loss of viability of 25–30% (Fig. 2b). The percentage of binucleate unbudded plus anucleate cells was similar (25–30%; Fig. 2a, d and Table 1), indicating that these cells may be inviable. *yeb1* mutants also have a mild mitotic defect (see below), which may contribute to this inviability.

We determined the relationship between this checkpoint and other checkpoint mechanisms. We used hydroxyurea and ultraviolet irradiation to induce the DNA-replication and -damage checkpoints and used nocodazole to induce the spindle-assembly checkpoint^{8,9}. As positive controls, we used a *mad1Δ* strain for the spindle-assembly checkpoint and *mec1* or *rad9* strains for the DNA checkpoints. During growth on solid media, hydroxyurea treatment and release from treatment or ultraviolet irradiation had no effect on wild-type and *yeb1* strains under conditions in which *mec1*

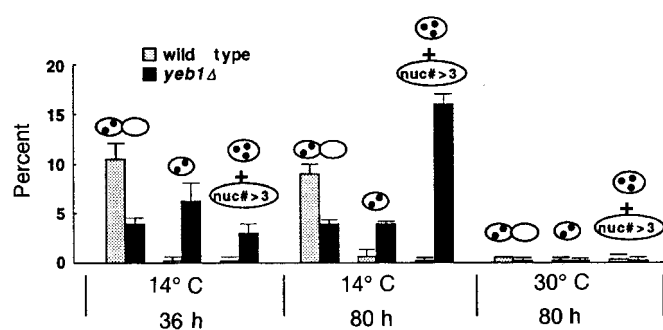


Figure 4 Effect of cold temperature on nuclear segregation. Wild-type and *yeb1Δ* cells grown to late log phase ($>10^7$ cells ml^{-1}) at 30°C were shifted to 14°C for 36 h or 80 h. Control cells were kept at 30°C. Data are shown only for abnormal cell forms: binucleate large-budded cells, binucleate unbudded cells, and multinucleate unbudded cells with three or more nuclei. More than 200 cells were counted for each sample. Error bars indicate s.e.m. Cells at 14°C divided slowly, reaching saturation during the time course of the experiment. Nuc, nucleus.

mutants did not grow. Nocodazole treatment and release from treatment had no effect on wild-type and *yeb1* strains under conditions in which *mad1Δ* mutants did not grow. Treatment of synchronized cells with hydroxyurea, ultraviolet light or nocodazole caused wild-type and *yeb1* cells to arrest with undivided nuclei ($<5\%$ cells with divided nuclei), 90 min after release from treatment with α -factor. No arrest occurred in a nocodazole-treated *mad1* strain (62% cells with divided nuclei), an ultraviolet-irradiated *rad9* strain (73% cells with divided nuclei), and a hydroxyurea-treated *mecl* strain (63% cells with divided nuclei). Therefore, the *yeb1* mutant has intact spindle-assembly and DNA-replication and -damage checkpoints.

We investigated how EB1 functions *in vivo*. APC binds to microtubules *in vitro* and *in vivo*^{10,11}. EB1 localizes to microtubules *in vivo*, in both budding and fission yeast^{12,13}. EB1 functions in the microtubule cytoskeleton of budding and fission yeast, judging from phenotypes resulting from loss of function and overexpression of EB1^{12,13}. An asynchronous population of *yeb1Δ* cells showed an increased number of large-budded cells with monoastal spindles. To investigate further, we examined the kinetics of spindle assembly and elongation. Formation of the short pre-anaphase bipolar spindle was delayed in *yeb1Δ* cells by ~ 50 min (Fig. 3a). The overall doubling time also differed by 50 min, with 210 min for *yeb1Δ* cells versus 160 min for wild-type cells. Elongation of the spindle from short to long was normal in the *yeb1Δ* mutant (Fig. 3b). The fine structure of the spindle and spindle pole bodies of *yeb1Δ* cells was unremarkable by thin-section electron microscopy (data not shown). Thus the *yeb1Δ* mutant has a mild defect in its rate of progression through the early phase of spindle assembly.

Cold temperature often aggravates mitosis and nuclear-migration defects^{2,14,15}, presumably because microtubules are unstable in the cold. *yeb1Δ* mutants grew poorly at 14°C, so we proposed that cold temperature induced the EB1 checkpoint through abnormal spindle orientation due to reduced microtubule stability. Among wild-type cells grown at 14°C for 36 h, $\sim 10\%$ were large-budded with both nuclei in the mother (Fig. 4); this defect is similar to that seen in dynein mutants at 30°C. Binucleate unbudded cells were rarely seen. *yeb1Δ* cells grown at 14°C for 36 h showed binucleate and multinucleate unbudded cells, in addition to large-budded binucleate cells. In *yeb1Δ* cells grown at 14°C for 80 h, numbers of multinucleate cells increased dramatically. The nuclear-migration defect in wild-type cells at 14°C and the multinucleate cells seen in *yeb1Δ* cells indicate that the EB1 checkpoint may be activated by the cold. Yeast in the wild experience cold temperatures, which may be the physiological activator of the EB1 checkpoint.

We tested whether the cold sensitivity of *yeb1* mutants was due to enhancement of the mitotic phenotype. The percentage of large-budded cells with a monoastal spindle, which reflects the mitotic phenotype of *yeb1* mutants, was 7% in *yeb1* cells at 30°C, 4% in *yeb1* cells at 14°C, and 1% in wild-type cells at either temperature. Therefore, the *yeb1* mitotic phenotype was not enhanced in the cold.

We have provided evidence for the existence of a new cell-cycle checkpoint that occurs before cytokinesis. This checkpoint is important for cells to maintain neutral ploidy under conditions that induce mitotic spindle misorientation, including cold temperature. As EB1 localizes to spindle microtubules and as the loss of EB1 slows spindle assembly, EB1 may be a necessary component of the checkpoint's sensor mechanism. After sensing the misorientated spindle, the EB1 checkpoint may activate a signal to delay the cell cycle. The phosphatidylinositol-4-OH kinase Stt4 may be an element of this signalling pathway. As aneuploidy is often observed in cancer cells, human EB1 may be necessary for APC's tumour-suppressor function through this checkpoint. Yeast do not have obvious sequence homologues of APC, but the function of EB1 in yeasts and humans may be similar because human EB1 rescues the *Schizosaccharomyces pombe* EB1 mutant¹². This checkpoint may also function in development when the orientation of the mitotic spindle is important for determination of cell fates. □

Methods

Strains. Strains were derived from YCH125:MATa *ade2 ade3 ura3 leu2 trp1 LYS2* (YJC1428) and YCH128:MATa *ade2 ade3 ura3 leu2 TRP1 lys2* (YJC1429) cells, donated by C. Hardy. Media and α -factor were prepared and used as described². Nocodazole (Sigma, St Louis, MO) and benomyl (DuPont, Wilmington, DE) at 15 mg ml^{-1} in dimethylsulphoxide were diluted 1,000-fold into media. 2 M hydroxyurea in water was diluted 20-fold into media.

Mutant isolation. YJC1259:MATa *ade2 ade3 ura3 leu2 TRP1 lys2 act5Δ::HIS3* [CEN ACT5 URA3 ADE3] and YJC1355:MATa *ade2 ade3 ura3 leu2 trp1 LYS2 act5Δ::HIS3* [CEN ACT5 URA3 ADE3] strains were treated with the mutagen ethyl methanesulphonate to 30% viability, and plated at 25°C for 4–6 days until red colonies appeared. Solid red colonies and red colonies with small white sectors were tested for growth on 5-fluoro-orotic acid. An ACT5 plasmid shuffle assay tested for specificity. Mutants were defined as recessive or dominant and placed into complementation groups.

Gene cloning and disruption. Mutants were transformed with a genomic library (ATCC 77162) and incubated at 25°C until sectoring colonies were identified. Polymerase chain reaction (PCR) was used to exclude ACT5 plasmids. Insert ends were sequenced and identified in the genome. The overlapping regions of different clones identified candidate open-reading frames (ORFs). Subcloning identified one ORF as necessary and sufficient for rescue. Linkage analysis showed that the rescuing DNA fragment was tightly linked to the original mutation.

Sequence analysis, including our analysis here, confirmed the high level of similarity between *S. cerevisiae* and human EB1, and the existence of EB1 homologues in *S. pombe* and mouse^{12,13}. Database searches with an EB1 protein identified only other EB1 proteins as having high sequence similarity.

A PCR-based method² was used to generate an allele with a complete disruption, *yeb1Δ::HIS3*. Haploid mutants derived from several independent diploids were viable.

Double mutants. *yeb1Δ* was synthetic-lethal with *dhc1-Δ7* (from K. Bloom) on the basis of tetrad dissection of 32 asci. All predicted double-mutant haploids were inviable. Similar results were obtained for *yeb1Δ* and *jnm1-Δ3* (provided by K. Tatchell) with 21 asci, for *stt4-7* and *dhc1-Δ7* with 20 asci, and for *stt4-7* and *jnm1-Δ3* with 19 asci.

Temperature-sensitive act5. A temperature-sensitive allele of *act5* was constructed with the temperature-sensitive degron system¹⁶. Temperature-sensitive *act5*, plasmid-borne or integrated, conferred temperature sensitivity to growth of three different synthetic-lethal mutants from the screen.

Movie assay. Microtubule and nuclear positions were determined with fusions of green fluorescent protein (GFP) to tubulin and to the SV40 nuclear-localization signal (NLS), respectively. GFP-TUB1 was from A. Straight¹⁷.

GFP-SV40-NLS was from J. Haseloff. These strains were also used in cell-cycle-delay assays with α -factor-mediated synchronization.

Cells were grown in synthetic media lacking methionine for 2 h to induce GFP-TUB1. During the second hour, the temperature was raised to 37°C to inactivate temperature-sensitive *act5*. Cells were placed on agarose with non-fluorescent growth medium supplemented with β -alanine (0.5 mg/l⁻¹), thiamine-HCl (0.2 mg/l⁻¹), biotin (2 μ g l⁻¹), calcium pantothenate (0.4 mg/l⁻¹) and inositol (2 mg/l⁻¹), and time-lapse movies were made¹⁸. At 4-min intervals, a focal series of bright-field and fluorescence images was collected. At cell separation, the daughter shifted its position relative to the mother in the bright-field image. We performed five experiments with temperature-sensitive-*act5* YEB1 cells and seven with temperature-sensitive-*act5* yeb1 Δ cells.

Synchronized cell assay. Temperature-sensitive-*act5* cells grown at 28°C were synchronized with α -factor alone or with α -factor followed by hydroxyurea¹⁹, shifted to 37°C for 20 minutes, and released from treatment at 37°C. α -factor was readed either after 30 min or immediately. Samples were fixed and stained with 4,6-diamidino-2-phenylindole. More than 200 cells were counted for each time point. Viability was determined by plating on YPD medium. Immunofluorescence using anti-tubulin antibodies or GFP-TUB1 confirmed that mitotic spindles were intact in the large-budded cells with both nuclei in the mother.

Cell forms as described in Table 1 were counted at each time point. With use of hydroxyurea, unbudded uninucleate cells at $t = 0$ (8–11%) were not synchronized and therefore assumed not to participate in the experiment. This percentage was subtracted from values obtained at subsequent time points. As one normal mitosis produces two unbudded uninucleate cells, the number of such cells was divided by two to calculate number of mitoses.

Received 27 February; accepted 25 March 1998.

1. Yeh, E. *et al.* Spindle dynamics and cell cycle regulation of dynein in the budding yeast, *Saccharomyces cerevisiae*. *J. Cell Biol.* **130**, 687–700 (1995).

2. Muhua, L., Karpova, T. S. & Cooper, J. A. A yeast actin-related protein homologous to that in vertebrate dynein complex is important for spindle orientation and nuclear migration. *Cell* **78**, 669–679 (1994).
3. McMillan, J. N. & Tatchell, K. The *JNM1* gene in the yeast *Saccharomyces cerevisiae* is required for nuclear migration and spindle orientation during the mitotic cell cycle. *J. Cell Biol.* **125**, 143–158 (1994).
4. Yoshida, S. *et al.* A novel gene, *STT4*, encodes a phosphatidylinositol 4-kinase in the *PKC1* protein kinase pathway of *Saccharomyces cerevisiae*. *J. Biol. Chem.* **269**, 1166–1172 (1994).
5. Su, L. K. *et al.* APC binds to the novel protein EB1. *Cancer Res.* **55**, 2972–2977 (1995).
6. Kinzler, K. W. & Vogelstein, B. Lessons from hereditary colorectal cancer. *Cell* **87**, 159–170 (1996).
7. Hartwell, L. H. & Weinert, T. A. Checkpoints: controls that ensure the order of cell cycle events. *Science* **246**, 629–634 (1989).
8. Elledge, S. J. Cell cycle checkpoints: preventing an identity crisis. *Science* **274**, 1664–1672 (1996).
9. Murray, A. W. The genetics of cell cycle checkpoints. *Curr. Opin. Genet. Dev.* **5**, 5–11 (1995).
10. Munemitsu, S. *et al.* The APC gene product associates with microtubules *in vivo* and promotes their assembly *in vitro*. *Cancer Res.* **54**, 3676–3681 (1994).
11. Smith, K. J. *et al.* Wild-type but not mutant APC associates with the microtubule cytoskeleton. *Cancer Res.* **54**, 3672–3675 (1994).
12. Beinhauer, J. D. *et al.* Mal3, the fission yeast homologue of the human APC-interacting protein EB-1 is required for microtubule integrity and the maintenance of cell form. *J. Cell Biol.* **139**, 717–728 (1997).
13. Schwartz, K., Richards, K. & Botstein, D. *BIM1* encodes a microtubule-binding protein in yeast. *Mol. Biol. Cell* **8**, 2677–2691 (1997).
14. Sullivan, D. S. & Huffaker, T. C. Astral microtubules are not required for anaphase B in *Saccharomyces cerevisiae*. *J. Cell Biol.* **119**, 379–388 (1992).
15. Li, Y. Y. *et al.* Disruption of mitotic spindle orientation in a yeast dynein mutant. *Proc. Natl Acad. Sci. USA* **90**, 10096–10100 (1993).
16. Dohmen, R. J., Wu, P. & Varshavsky, A. Heat-inducible degenon: a method for constructing temperature sensitive mutants. *Science* **263**, 1273–1276 (1994).
17. Straight, A. F. *et al.* Mitosis in living budding yeast—anaphase A but no metaphase plate. *Science* **277**, 574–578 (1997).
18. Waddle, J. A. *et al.* Movement of cortical actin patches in yeast. *J. Cell Biol.* **132**, 861–870 (1996).
19. Weiss, E. & Winey, M. The *Saccharomyces cerevisiae* spindle pole body duplication gene *MPS1* is part of a mitotic checkpoint. *J. Cell Biol.* **132**, 111–123 (1996).

Acknowledgements. We thank C. Hardy, J. Haseloff, M. Johnston, A. Murray, A. Straight, D. Lew, M. A. Hoyt and M. Rose for strains and plasmids; and T. Karpova, K. Blumer, C. Hardy, R. Heil-Chapdelaine, A. Murray, S. Wente and M. Winey for advice and discussion. This work was supported by grants from the NIH, N.R.A. is a recipient of a Natural Sciences and Engineering Research Council of Canada postdoctoral fellowship. J.A.C. is an established investigator of the American Heart Association.

Correspondence and requests for materials should be addressed to J.A.C. (e-mail: jcooper@cellbio.wustl.edu).

KNOW YOUR COPY RIGHTS RESPECT OURS

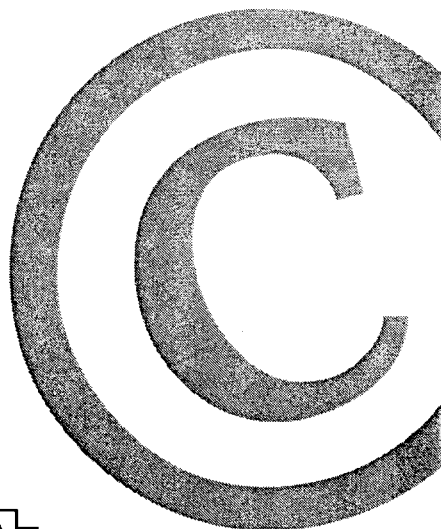
The publication you are reading is protected by copyright law. This means that the publisher could take you and your employer to court and claim heavy legal damages if you make unauthorised photocopies from these pages.

Photocopying copyright material without permission is no different from stealing a magazine from a newsagent, only it doesn't seem like theft.

The Copyright Licensing Agency (CLA) is an organisation which issues licences to bring photocopying within the law. It has designed licensing services to cover all kinds of special needs in business, education, and government.

If you take photocopies from books, magazines and periodicals at work your employer should be licensed with CLA.

Make sure you are protected by a photocopying licence.



The Copyright Licensing Agency Limited
90 Tottenham Court Road, London W1P 0LP
Telephone: 0171 436 5931
Fax: 0171 436 3986

Investigating the alternatives

Fewer than half of doctoral graduates head for a career in academic research. But the range of valuable skills learned is increasingly being put to good use elsewhere, with a growth in opportunities in industry and business.

Helen Phillips

Only a minority of today's PhD graduates can realistically expect long-term careers in university research. This has been acknowledged by research organizations including the US National Academy of Sciences (NAS) and National Science Foundation (NSF), and by the UK government. In the United States, just one-third of those receiving doctorates are expected to enter the academic tenure system. But, despite these realities, hiring in other areas has been vigorous enough to keep the overall level of unemployment among PhDs relatively low. And students are adjusting their expectations accordingly and considering established and emerging career paths for scientists outside the area of academic research.

The US Committee on Science, Engineering and Public Policy (COSEPUP), which brings together the NAS, National Academy of Engineering, and the Institute of Medicine, reports that "more than half of new graduates with PhDs—and much more than half in some fields, such as chemistry and engineering—now find work in non-academic settings". This fraction has been growing steadily for two decades.

But this movement does not reflect an abandonment of the skills developed while working for the degree, or even necessarily a movement away from research. Mark Regets, a labour force analyst at the NSF, says: "People to a great extent end up using their degrees. The rate of involuntary employment outside the field of study is low." The UK Office of Science and Technology also reports that most PhD graduates found their studies were helpful and related to the careers they were now following.

And there are certainly attractions to the alternative career paths. The NSF has reported that median salaries in 1995 for recent PhDs were highest in the private, non-education sector (\$56,000) and lowest for post-doctoral researchers (\$28,000).

Alternative destinations

So just where are these attractive alternatives to be found? The US Bureau of Labor Statistics predicts that employment in science and engineering occupations will increase to the year 2006 at more than three times the rate for all occupations. Three-quarters of this increase will be in computer-related occupations (see page 497) and there will be significant growth in health-care services.

Nick Jagger, from the UK Institute of Employment Studies, says that currently the

commonest destination for PhD graduates is industrial research and development (R&D). But there are other increasingly common destinations. "Some enter consultancies and others adopt a marketing role, selling into technical markets," he says. (See Fig. 1.)

COSEPUP highlights investment banking, medicine, government, computing, small businesses, law, education and science journalism as important non-traditional sectors of employment. Patrick Mulvey of the American Institute of Physics says: "Permanent academic positions are not in abundance, but industry is absorbing physics doctorates." The institute is encouraging students to diversify. "These are not necessarily new career paths," says Mulvey, "but just not previously highly publicized ones." Mulvey reports that 69% of potentially permanent employment for physicists was taken in the

industrial sector. Other than industrial R&D, the main destinations for physicists seem to be computer-related fields and the financial sector (see page 496).

The Federation of American Societies for Experimental Biology (FASEB) reports a similar increase in the industrial employment of biomedical PhDs: "Jobs in industry have increased and, in the future, might surpass academic jobs as the most prevalent form of employment for US biomedical scientists." The rise is particularly evident in the pharmaceutical and biotechnology industries, and in the relatively new field of bioinformatics (see *Nature* 389, 417; 1997).

FASEB reports that, with the increase in clinical, agricultural, forensic and other applications of biological research, the need for biomedical PhDs in affiliated industrial sectors will increase, as will the demand from

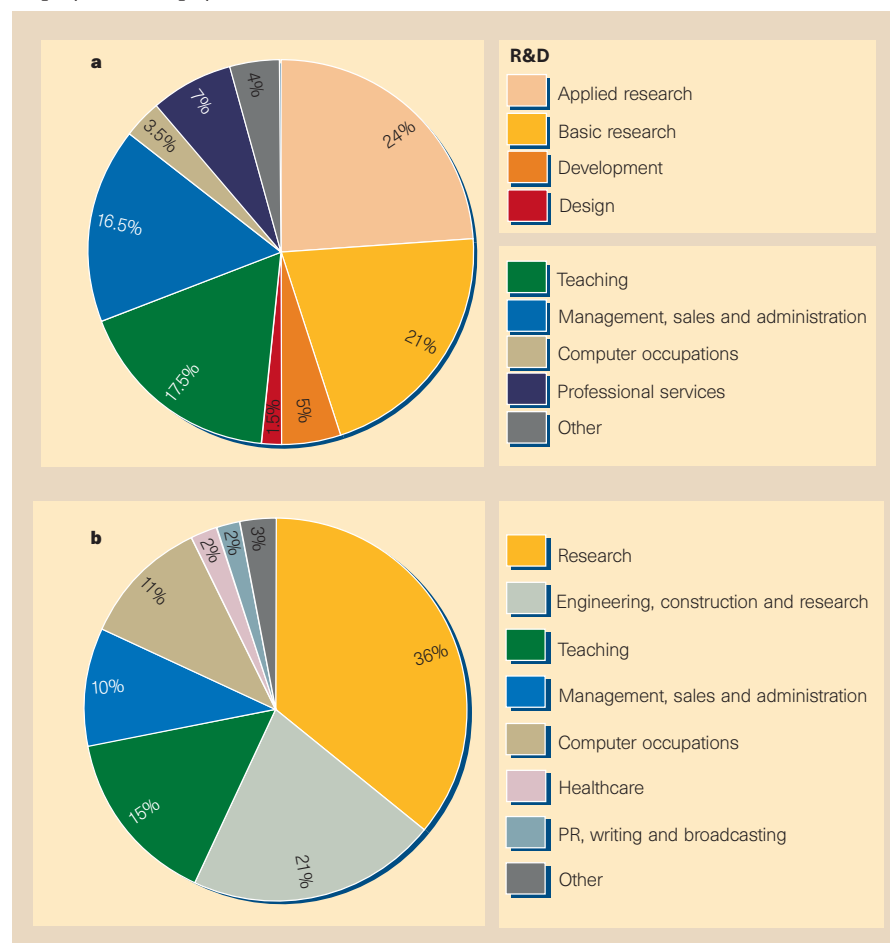


Figure 1 a, Sector of employment of doctoral scientists taken from the NSF Division of Science Resources Studies report, *Characteristics of doctoral scientists and engineers in the United States*, compiled from the Survey of Doctoral Recipients from 1995. This surveys persons under the age of 76 who hold doctorates in science and engineering from US institutions. b, Results from the UK Office of Science and Technology survey of postgraduates funded by the research councils, roughly 10 years after graduation.

other occupations such as patent law, public policy and science writing.

But in many European countries a doctoral degree is still considered a uniquely academic qualification. Gerhard Neuweiler, former president of Germany's science council, the Wissenschaftsrat, says that doctoral graduates from German universities are encouraged to take academic posts abroad, often with German funding, rather than pursuing alternative careers.

However, in many countries, the gradual acknowledgement of the trend towards non-academic employment has led policy-makers to emphasize the need to train postgraduates in more general skills. The COSEPUP report says: "Graduate education should prepare students for an increasingly interdisciplinary, collaborative and global job market." Similarly, the UK government says "greater attention will be given to the relevance of post-graduate training for all careers".

The consensus is that industry is looking for highly trained individuals with a good technical grounding, good writing skills, an understanding of statistics, good interpersonal skills, and software and mathematical skills. Those who make good negotiators and team leaders, and who can work independently, are also favoured.

Making the break

COSEPUP has recommended that students should be given more advice by universities about the broader range of career choices available. This should come as welcome news to many, including the American Institute of Physics, which reports that people do not

necessarily know where to look to find employment outside the academic world. Many university careers services are tailored more towards first-degree graduates.

Neuweiler paints a bleaker picture of the careers advice available in Germany. "There is no advice, nothing," he says. But there is encouragement from some quarters. The UK research councils fund a Graduate Schools programme to help students think about the skills that would be useful in many alternative careers.

Increasingly, postgraduate students are setting up Internet discussion groups and self-help forums. Kate Hogan and other graduate students at the University of North Carolina have collaborated with their department chairman and the director of graduate studies to create an 'alternative careers forum' for students and postdocs. Hogan says: "Most of us have spent the vast majority of our time in the lab and have a bad case of tunnel vision." Many see their options as staying in academic life or joining industry but, says Hogan, "there are far more opportunities than that". Research on the Internet bears this out — there are many websites on career alternatives for scientists (see Table 1).

The forum's popularity has shown that students are looking at what else the future may hold, and with each session the list of options that people would like to have represented grows, says Hogan. "We also have plans to educate the faculty about emerging opportunities for students."

Other trends include the development of postdoc societies, such as those that originated at Johns Hopkins University and the Uni-

versity of California at San Francisco. Such organizations have improved the available information for making career decisions, and helped to create larger forums such as those at professional society meetings.

Help is available but a career is a 'do-it-yourself' project, as Stephen Rosen, chairman of the Science and Technology Advisory Board and director of Scientific Career Transitions, in New York, points out. His general advice to those who are wondering where to start looking for an alternative career is: "All the subsidiary skills one has to develop as a scientist are potential career directions. Almost any destination is possible from a science position."

Helen Phillips is a science writer for Nature, based in London. e-mail: h.phillips@nature.com

Reading recommendations

Robbins-Roth, C. (ed.) *Alternative careers in science* (Academic, San Diego, 1998).

Rosen, S. & Paul, C. *Career renewal: tools for scientists and technical professionals* (Academic, San Diego, 1998).

COSEPUP *Careers in science and engineering: a student planning guide to grad school and beyond* (National Academy Press, Washington DC, 1996).

The fast track, by Mariam Naficy, covers management consulting, investment banking and securities trading for MBAs and undergraduates. An interview with the author discusses options for those with higher technical training.

Careers in science and engineering, from the NAS: www.nap.edu/readingroom/books/careers

Legal recourse

Potter Wickware

Career opportunities in environmental, regulatory and patent law are available for people with strong science backgrounds. Although some scientific disciplines, and the legal profession itself, suffer from overpopulation at present, those who combine science and law can create satisfying career alternatives and meet needs in industrial and governmental organizations.

Susan Poulter, who teaches environmental law at the University of Utah, says: "There are many legal clients who are eager to assess the risk of their operations in the light of environmental regulations. But the pesticide, clean air and water statutes are complex, and interpreting them requires good scientific understanding on top of legal knowledge." Poulter, who holds a PhD in organic chemistry from the University of California, Berkeley, and taught at Utah before embarking on her legal education, foresees environmental law work remaining abundant in the future, even if it is not today quite the rising star it was a few years ago.

Scientific training can be a direct preparation for legal work, but the link is not always so clear. Richard Meserve of Covington & Burling, a large law firm in Washington, DC, specialises in toxic torts and says that his doctoral training at Stanford in the properties of paramagnets has never inter-

Table 1 Useful careers resources on the web

General	
Nature	URLs: http://www.nature.com/
CareerPath.com	www.careerpath.com
Science's Next Wave	www.nextwave.org
Biomednet	biomednet.com/jobs
Nature/Scientific American new careers site	www.scitechjobs.com
Organizations and Societies	
National Academy of Sciences' Career Planning Center	www2.nas.edu/cpc
Hobson graduate employment and training home page	www.get.co.uk/index.htm
US News Online Career sites	www.usnews.com/usnews/edu/beyond/bcclinks.htm
Specific subject areas	
Management consulting	www.cob.ohio-state.edu/~fin/jobs/mco/mco.html
MIT's Sloan School of Management	web.mit.edu/sloan/www/
Patents.com (Oppedahl & Larson)	www.patents.com
American Institute of Physics	www.aip.org/careers/
Harvard Forum on creating careers in physics	phys3.harvard.edu/careers/
Yale's Alternative Careers in Bioscience Seminars	www.mbb.yale.edu/acb/seminar.html
American Chemical Society	www.acs.org/careers/welcome.htm
American Geological Institute	agi.umd.edu/agi/careers.html
Minerals and Materials Society	www.crc4mse.org
American Geophysical Union	www.agu.org
Graduate groups	
National Postgraduate Committee	www.npc.org.uk
Gradlink	www.gradlink.edu.au



Estreicher: key role in advising designers.

sected directly with his legal work. But "sound scientific training is an excellent foundation for a career in many legal areas, even if never directly applied", because the scientist's ability to analyse phenomena mirrors the lawyer's ability to sort out the rules governing the resolution of a legal matter.

The advantage of science applied to law is manifest in environmental cases that hinge on scientific interpretations, says Herbert Estreicher, who specializes in chemical regulation at Covington & Burling.

Legal knowledge can also improve scientific work, particularly where research is directed at developing products. Estreicher, who has a PhD in chemistry from Harvard, and went to law school at New York University, says that the biotechnology, chemical and consumer product industries are subject to many statutes that product designers are wise to take account of early in development cycles. A scientist-lawyer can help the company to anticipate risks and maximize the chances of market approval. "Increasingly, products are required to be safe when they are new and benign when they are worn out. Efficacy and cost are well known design constraints, but a third, underappreciated one is regulatory approval," he says.

Despite the opportunities available to scientist-lawyers, the hard reality is that



Chadwick: from biophysics to patent law.

today there is an abundance not only of PhDs in geology, physics and biomedicine, but of lawyers as well, and the job market can be tough for new entrants.

Robin Chadwick, who received a PhD in molecular biophysics from Yale and now practises patent law at Kenyon & Kenyon in New York, says her field is generally a good choice for scientists, and that patent lawyers are in demand now, particularly in chemical and biotechnology-related disciplines. The US Patent and Trademark Office (PTO) also suffers from a chronic shortage of patent examiners in these areas, a need that is bound to become aggravated as the rate of patent activity continues to rise.

The scientist interested in patent work need not necessarily spend the three years typically required to get a juris doctor law degree to get work, points out Ric Snead. He has a BS in zoology and an MS in biochemistry from Clemson University, and is now a

patent expert at Dialog, an online information company in Palo Alto, California. A patent agent drafts patent applications and deals administratively with the PTO, but unlike a patent attorney cannot argue cases or give legal advice to clients. Despite the limitations of agent status, Snead has found the qualification useful in his career.

To qualify for the examination to become a patent agent or attorney, which is given once a year by the PTO, a minimum of 40 hours of undergraduate science or engineering course work is needed. To pass the exam it's probably advisable to take a cram course, says Snead. He adds that to become an agent one should realistically allow a year, preferably in a legal environment, perhaps doing an internship in preparation.

But Poulter cautions that the investment required to enter the legal profession, whether in patent, environmental or regulatory law, is substantial, and the effort should only be undertaken after weighing the benefits and costs as carefully as when contemplating entering a PhD programme. □

Potter Wickware is a science writer in Oakland, California, USA.

e-mail: wick@netcom.com

Reinventing yourself

Robert Goriss

Seeking employment in the patent business is not merely appealing because of the earnings potential. When compared with starting academic salaries or the pay of a postdoc, it is well paid work. But it is also scientifically challenging for graduates who do not want to work at the bench, or researchers looking for a change.

In Europe, scientists can enter the patent business as patent examiners in national patent offices or the European Patent Office (EPO), as patent agents or as patent attorneys. The meaning of these terms varies slightly around the world. Employees in patent departments of commercial companies are sometimes called patent officers.

To become a patent examiner requires up to two years on-the-job training, particularly in relevant aspects of law. One important aspect of the job involves finding out if a similar patent claim has already been filed somewhere. This requires computer training because patents and applications are now routinely stored in databases. Another important aspect is deciding whether the patent claim is sufficiently novel and inventive to warrant a patent. Visits to technology exhibitions, research institutes and companies are essential to keep examiners up to date.

Germany and the United Kingdom have Europe's biggest national patent offices. The UK office takes on graduates in mathematics, physics, biology, chemistry and engi-

neering with first- or second-class honours degrees. "New employees are trained on real patent applications in the first year, when they are intensively coached by two training officers," says Peter Back, head of examiner recruitment. "In the second year the level of coaching is reduced, and more independent work is demanded." On average, it takes three to five years to become really independent, he says. Last November, 20 examiners were hired, and next November a further 30 will be taken on.

The German patent office, in Munich, employs around 600 patent examiners, and last year 10 new examiners were taken on. It does not hire fresh graduates, but demands at least five years research or development experience in industry or the academic community. Examiners usually start to work independently after 18 months of training. Beginners earn around DM4,500 a month (US\$2,500). Experienced examiners may earn DM7,500 a month (US\$4,200).

"To summarize the results of a search and examination in a new technological area is a creative and interesting occupation," says Jürgen Mume, head of the German Federation of Civil Servants and Judges in Patenting, and an examiner himself. "The examiner has sole responsibility for the acceptance or refusal of an application and has an almost judge-like position."

Half of the EPO's 3,700 employees are patent examiners. As the EPO comprises 19 European countries, employees must be conversant with the EPO's three official languages, English, French and German. Candidates must have a university degree in science or technology. New recruits are given one year's training on the job. Remuneration, more or less tax-free, is generally up to twice as high as that at national patent offices.

Despite good pay and benefits, the EPO finds it hard to attract appropriately qualified people. This year the EPO will hire 140 patent examiners, but a high proportion of them will be in the telecommunications division, which is one of the most rapidly developing areas. The EPO has the problem of needing to employ an appropriate mix of nationalities. "We find ourselves desperately looking for Finnish engineers, for example," says Rainer Osterwalder, an EPO spokesman. "Scandinavia's leading telecom-

Table 2 Further information on careers in patenting

A list of and links to the world's national patent offices can be found on the Internet:

http://www.deutsches-patentamt.de/recherche/rech_6.htm

European Patent Office:

<http://www.european-patent-office.org>

The UK Patent Office Search and Advisory Service is at Cardiff Road, Newport NP9 1RH.

Tel: +44 1633 811 010.

Chartered Institute of Patent Agents:

<http://www.cipa.org.uk>

munications industry is snapping them up after graduation, before we can get to them."

Patent examiners may go on to work in patent departments of companies or may become patent agents, a profession that requires training in drafting patents and patent laws, and an examination is necessary to be allowed to represent inventors. Experience as a patent examiner is not necessary to become a patent agent, but it helps.

In Britain, these exams are set once a year, in November, and consist of two parts, the first usually after a year's professional training, the second after two or three years. Candidates are usually trained while working as technical assistants to a patent agent in a firm of agents or in an industrial patent department. "Although it is not a requirement, we want candidates to have three years of professional training," says Michael Ralph, spokesman of the Chartered Institute of Patent Agents.

Patent agents as well as solicitors in the United Kingdom are allowed to call themselves patent attorneys and to act as such without any further examination. In Germany the *Patentanwalt* corresponds to the British patent agent. A two-year training at a *Patentanwalt* office is followed by a year either at the German Patent Office or the Federal Patent Court in Munich. Examinations are set by the German Patent Office.

Any patent agent is eligible to sit EPO exams to qualify as an European Patent Attorney without further training. In fact many patent agents in the United Kingdom are also registered with the EPO as European Patent Attorneys. The job of a European Patent Attorney, not to be confused with patent attorney, is to consult inventors and companies in patenting affairs and to draft applications for a European patent. □

Robert Goriss can be contacted through the Munich office of Nature.

Taking your skills all the way to the bank

Diane Gershon

There used to be a feeling among academics that people with PhDs in maths or physics went into finance because they couldn't find a job in the academic world, and that somehow these jobs were 'second best', says Peter Knight, a physics professor at Imperial College, London. But when he talked to some of these people, Knight found that they were using the same sorts of professional skills as they would in an academic job but in a very practical setting, and being paid a lot more money to do it. "I don't regard it as a failure... all we've done is generate someone with marketable skills," he says.

Indeed it appears that investment banks

Table 3 **Financial information on the web**

Company	http://
J.P. Morgan	www.jpmorgan.com/
Goldman Sachs	www.goldmansachs.com/
Bankers Trust	www.trustnet.co.uk/
Nikko Securities	www.nikko.co.jp/index_e.html
Paine Webber	www.painewebber.com/
Investment Banking (Fisher College of Business)	www.cob.ohio-state.edu/dept/fin/jobs/ib.htm
Engineering, Physics and Science Research Council graduate schools	www.epsrc.ac.uk/progs/schemes/grad-hal.htm

are targeting people at the PhD level who have the right mix of mathematical ability and computational know-how. A 1997 survey by the UK Engineering and Physical Sciences Research Council (EPSRC), which looked at where people were employed one year after obtaining their PhDs, found that not only was there little or no unemployment among new physics PhDs, but also that the financial sector was the major employer in some specialist areas. It gave jobs to 48% of those with PhDs in mathematical and statistical physics, for example.

For scientists with a strong background in applied maths and with considerable computing experience, there are two potential avenues of employment in the quantitative and modelling-intensive area of derivatives: quantitative analysis or mathematical programming and code development. (Jobs are now also opening up for ex-scientists in 'front-office' trading positions after they have gained several years experience on the quantitative side.) Posts in quantitative analysis require a good knowledge of higher stochastic calculus and statistics but also include a substantial amount of computing (C++ and the like). "It is sometimes astonishing how little practical computing experience scientists bring with them," says Peter Jaekel, a quantitative analyst, or 'quant', in the risk management department of Nikko Securities Global Holdings in London.

Jaekel, who joined Nikko a little over a year ago as part of its newly formed global risk management group, warns that it is quite difficult to land a plum job. It takes perseverance and commitment and the interview process can be gruelling. Financial institutions can also afford to cherry-pick the best and most committed applicants, and are generally looking for people who are able to learn quickly, who pay attention to detail (as mistakes can be costly in this field), who like a challenge and don't go running when the first problem arises. "We receive only a pre-screened selection of job applications as they are usually provided by headhunters. Out of those, we only invite the top 20–40 per cent for an interview," says Jaekel. Most initial appointments are made through recruitment agencies (some of which can be found on the Web), or sometimes by word-of-mouth recommendations. Once inside the field, personal contacts become increasingly important for career advancement, and the

dogged pursuit and poaching of good people is a way of life.

Jaekel says that a committed individual can usually start in the investment business without further training but the standard text, *Options, Futures and Other Derivatives* by John Hull (Prentice Hall), is a must read. Also, read the *Financial Times* or *Wall Street Journal* and keep abreast of financial indicators and economic trends.

The work itself is "very academically structured, only geared towards financial application," says Jaekel. He regularly puts in a 60-hour week and spent his first two months at Nikko studying the purest mathematical number theory that he had ever seen. For him, the job has meant more stimulus, more challenge and higher potential rewards. He says that no PhD scientist starts in the financial sector on less than £25,000 (US\$40,700) a year, and most earn £30,000–35,000.

From the outside, the world of high finance seems a closed and mysterious place — particularly given the dizzying amounts of money that change hands. But the boom in the derivatives market in recent years has opened up a new level of mathematical (and computing) complexity for investment houses, such that mathematicians and physicists are finding that, even though they may have no market experience, they have bankable skills and a lucrative non-traditional career option. But it may not be everyone's cup of tea. It's not a job for the faint-hearted. When Jaekel was contemplating such a career move, someone advised him: "Don't do it. It eats your soul." Jaekel's response to that today is: "Well it hasn't, and I am a lot happier now." □

Diane Gershon is assistant editor, new technology, at Nature Medicine.

Further reading

Stix, G. *A calculus of risk*. Sci. Am. 92–97 (May 1998).

The write stuff

Helen Phillips

Science journalism is an enticing option for scientists and students who feel that research does not provide them with the breadth or immediacy they would like. But how easy is it to make the transition from bench-top to

lap-top? Unsurprisingly, opportunities in the more glamorous broadcast media are rare and highly sought after. But consider science writing, journalism, editing, broadcasting, technical and medical writing together, and the prospects look much more favorable.

There are two common warnings from those in the business. First, writing science in simple language is not simple to do. "It's no easier than getting a PhD," says John Wilkes, director of the graduate programme in science communication at the University of California at Santa Cruz. Second, writing a thesis or scientific papers is no training for communicating to a popular audience.

Warnings aside, science journalism is an extremely rewarding career, with variety, immediacy and the thrill of staying in touch with the latest research.

Breaking in

There is no set career structure for science journalism but the first step on the ladder seems to be the crucial stage. "Getting the first opportunity is the tricky one," says Joe Palca of US National Public Radio.

To find that first break, one option is to take one of the growing number of training courses in journalism and science communication (see Table 4) — although there is strong competition for places. And more and more of the applicants have PhDs. But scientific credentials are not enough — Wilkes is looking for those who are "already first-rate writers" for his course at Santa Cruz. Courses can encourage new ways of thinking about science, and provide training on writing structure, experience of different media formats and a safe place to practise. Phyllida Brown, freelance writer and consultant for *New Scientist* magazine, says courses can help to build confidence and allow students to develop an understanding of the way the media industry works — such as how to work within a brief and with an editor.

For others, the most helpful component of a course is a work placement, frequently a stepping stone into employment. Another advantage of taking a course is that increasingly employers are looking to them to make

their task of selection easier, explains Nick Russell, director of the science communication course at Imperial College, London.

But courses are not the only way in. "Training helps but it is not absolutely essential," says Jon Turney, writer and senior lecturer in science communication at University College London: "I'd say look around university while you are still in it for workshops and the like." Michael Kenward, science writer and editorial consultant, echoes this: "If you want to write, then start doing it in your spare time as a scientist." He also suggests another option. "A traditional route is into a specialist title that needs your scientific knowledge," he says. "There you will learn some of the skills of journalism and editing."

The specialist journals route is also recommended by Giles Newton, an editor at the Wellcome Trust in London, who began his career at *Current Biology*. To gain a solid foundation in editing and writing skills, he advises, "find somewhere that will teach you how to copy-edit". Another way of making contacts, especially helpful in the broadcast media, is to apply for an internship or fellowship programme. Science journalism associations (see Table 4) should be able to advise about local job opportunities.

If the media limelight isn't your goal, there are positions as press officers, science writers in residence, and information officers at research agencies and universities, especially in the United States. Or try the Technical Communicators' Society website and those of the medical writers' associations for a different type of opportunity.

Chris Christopher, who teaches on the technical writing certification programme at the University of California at Berkeley, made the transition from an academic career to technical writing herself, and says she has "felt highly rewarded both intellectually and financially". There are "enormous opportunities for PhD scientists," she says.

Novartis, the pharmaceuticals company, has five medical writers in Britain, and in this industry too the numbers are rising.

One final question is, how well does a PhD prepare you for a career in science journalism? Richard Keeble, author of *The News-*

papers Handbook, says the newspaper industry remains sceptical about the values of an academic training. But at a publication such as *Nature*, most assistant editors have post-doctoral research experience. Probably the biggest growth area for scientists-cum-writers is in the second-level review journals, says Turney. Science training can help with understanding the jargon and the process of research, but one also has to be willing to work outside one's own field, to ask stupid questions and get to grips with deadlines.

Although job opportunities are not common, Paul Willis of the Australian Broadcasting Corporation — whose PhD was on Australian fossil crocodilians — says that they are "more common than positions for dead croc workers". □

Reading on writing

Blum, D. & Knudson, M. (eds) *A field guide for science writers* (Oxford University Press, 1997).

ABSW *So you want to be a science writer?* (booklet available from EUSJA website or ABSW).

Keeble, R. *The newspapers handbook* (Routledge, 1995) New edition due later this month.

The directory of science communication courses and programs in the United States, University of Wisconsin-Madison (1996). \$10.

Available from Professor Sharon Dunwoody, e-mail: dunwoody@facstaff.wisc.edu

Scientists have a head start in IT

James Porteous

These are heady days for anyone wanting to work in the multiplying layers of the information technology (IT) industry. While the computer revolution surges and organizations invest heavily to keep pace, the demand for employees with the basic wherewithal is remarkably high. By 'tooling-up' and choosing an approach matched to their skills, scientists of all backgrounds will be well placed to take advantage of the diverse and lucrative opportunities available across the industry.

As processors and software have become more accessible, scientific fields outside the theoretical computing-based disciplines have been quick to adopt information systems in research and teaching. Now, in addition to their analytical skills, scientists are collecting industry-relevant computer experience from their everyday roles, meaning that they can be marketable as IT professionals both inside and outside science.

John Simpson, senior careers adviser at Imperial College, London, says: "Last summer, in just four weeks, we were approached by 80 IT-related companies looking to take on scientists with minimal skills. And it's the same this year. Anyone with a basic knowledge of computing systems and reasonable numeracy is in hot demand."

Similarly, a report on "Occupational employment projections to 2006" in the

Table 4 **Journalism and media information on the Internet**

Euro. Union of Science Journalists' Associations	www.esf.org/eusja/index.htm
Association of British Science Writers (ABSW)	www.esf.org/eusja/writer.htm
National Association of Science Writers (USA)	www.nasw.org/
Australian Science Communicators	www.asap.unimelb.edu.au/asc/asc.html
Society for Technical Communication (USA)	www.public.iastate.edu/~stc_isu/
American Medical Writers' Association (USA)	www.amwa.org/
European Medical Writers' Association	www.emwa.org/

Course information

Australian Science Communications	www.asap.unimelb.edu.au/asc/courses.html (Users may have to open main site first, then type the rest of the address once in)
The Wellcome Trust	www.wellcome.ac.uk/wellcomegraphic/a2/index.html

Internships and fellowships

AAAS mass media fellowships	www.nextwave.org/cgi/content/full/1998/04/13/15
-----------------------------	--

November 1997 *Monthly Labour Review*, produced by the US Bureau of Labour Statistics, predicts that by 2006 the number of IT-related positions will more than double to 2.5 million in the United States.

As Simpson points out, there is a need for IT aptitude everywhere, not only in finance, consultancy, general business, and software or hardware companies, but also in pharmaceuticals, biotechnology and bioinformatics. The application of information systems technologies to all areas of industry means that there is a varied range of IT responsibilities available within organizations.

Although 'information technology' is a very broad term — covering any application of computing — IT positions tend to fall into two loose categories: programming roles, involving code writing, systems development and modelling; and operations roles, involving the installation, running and administration of systems, controlling the distribution of information, and managing teams and projects. Outside these responsibilities lie opportunities revolving around IT developments, such as content management, and marketing.

The juiciest salaries are being paid by the finance industry to dedicated specialists, generally from hard computing backgrounds, who are able to write and implement customized modelling systems. But, as Sverre Jarp, a senior systems coordinator at the European Laboratory for Particle Physics, CERN, points out, scientific research is becoming so sophisticated that scientists are evolving into IT all-rounders. They have the added bonus of being able to work well in teams and having good self-management skills. "Because particle physicists, for example, need to be highly specialized programmers, they are being poached by banks and consultancies where their modelling skills and backgrounds are in great demand," says Jarp.

Shell, Sun Microsystems, AOL and Andersen Consulting are interested in IT-specialized scientists, but they all place value on retraining the high-achieving, personable scientist with good general IT exposure and problem-solving, decision-making and communication skills. Reflecting this, the recruitment consultancy Michael Page in London says that team-working scientists with such backgrounds could fill the parallel demand for IT-related project managers.

With the current array of willing targets for your job applications, it is worth contemplating where your computer experience may be in demand, and how you could enhance it. Use the Internet to learn about the industry and build your experience; keep an eye on key job websites (such as www.topjobs.com, www.nature.com and www.nextwave.org) and relevant company home pages, mapping your experience to their needs and looking at where you might

need to develop new skills. The Internet holds a wealth of self-tutoring information on everything from website preparation to more complicated coding and hardware updates — a little tuning of the search engine is all you need to find it.

Courses, such as Imperial College's MSc in computing science, are being developed to help scientists become IT-competent, but don't be afraid to take smaller steps. A science background is an excellent entry to the fringes of IT in many organizations, where training is offered. This approach will offer you the perspective needed to choose a preferred layer of the industry. □

James Porteous is web production editor at Nature.
e-mail: j.porteous@nature.com

Banking on biotech

Diane Gershon

In the realm of biotechnology finance, people are everything. Biotech analysts are the people responsible for making rational buy-sell decisions on a company's stock — information that is channelled to institutional buyers who depend on analysts to place a 'value' on a company's research and product portfolio in terms of a market opportunity.

According to Stelios Papadopoulos, managing director and head of the health sciences investment banking division at Paine Webber in New York, there are about 60 biotech analysts on Wall Street in 40 or so firms. Two-thirds of them have PhDs or MDs, and the rest have Masters in Business Administration (MBA) or similar backgrounds. That's a big difference from 15 years ago, he says, when it was just "the few of us [competing] against an unfriendly system" that had little appreciation of this nascent industry and the need for analysts to cover it.

Despite the increase, Papadopoulos says that in some ways it is more difficult to enter the business now. Competition is intense as there are "so many more scientists sensitized to the opportunities on Wall Street".

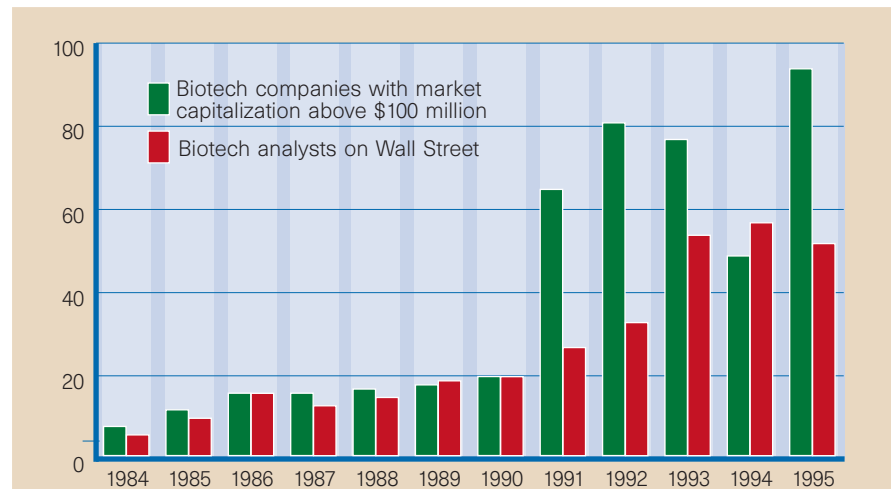
Papadopoulos got his PhD from New York University Medical Center in 1980. Around that time he was contemplating a career on Wall Street — a move that would allow him to marry his interest in science with business and, at the same time, offer considerably better earnings than the academic life. He was spurred on to pursue this goal after reading about Genentech's successful initial public offering of shares and promptly signed up to study for an MBA at business school by night while continuing with research by day.

In 1985, he finally landed a job as the first biotechnology analyst to be employed by securities firm Donaldson, Lufkin & Jenrette. He is now one of the most respected investment bankers in his field (investment bankers raise the large sums of cash needed to fuel this capital-intensive industry).

Papadopoulos says that what firms are looking for when they hire analysts is "intangibles — a certain sense that the person is capable of dealing with the complexity, the fuzziness, the ambiguous nature of Wall Street, and succeeding".

Although the science in biotech companies can be engaging, Papadopoulos says analysts must never lose sight of the fact that they are being paid to pick winners and, in doing so, to make money for other people. So many of the attributes that go to make a good analyst may not come easy to most scientists, who, like the athlete specializing in the 100-metre dash, are trained to focus narrowly on a specific area of science, says Papadopoulos. "By comparison, a job on Wall Street is the decathlon. You don't have to be outstanding at everything but you have to be fairly good at many different things."

So if you are personable, open, engaging, quick on your feet, able to change course mid-stream, and can learn to operate on the basis of scarce data, assessing risk and making decisions quickly, then a career as a biotech analyst may be for you. Energy and commitment are important factors as well: some analysts work 12- to 14-hour days and spend much time on the road. □



PAINE WEBBER